The Dissertation Committee for Gonzalo Enrique Espinoza Dávalos Certifies that this is the approved version of the following dissertation:

Large-Scale Statistical Analysis of NLDAS Variables and Hydrologic Web Applications

Committee:

David R. Maidment, Supervisor

Daene C. McKinney

Paola Passalacqua

Ben R. Hodges

Zong-Liang Yang

Large-Scale Statistical Analysis of NLDAS Variables and Hydrologic Web Applications

by

Gonzalo Enrique Espinoza Dávalos, B.E.; M.S.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2016

Dedication

I would like to dedicate this dissertation to my parents Berenice and Gonzalo for empowering me to pursue and achieve my academic and professional goals, to my siblings Miriam and Antonio for their unconditional friendship, and to my niece Regina, my nephew Mateo, and my brother-in-law Adrian.

Acknowledgements

I would like to thank Dr. Maidment for his committed role as an advisor, for his support, and for all the experience that I have acquired as part of his research group. I would like to thank Dr. Arctur for his guidance and collaboration, Dr. Teng for sharing his vision and for providing the tools and data access to develop my research, and Dr. Whiteaker for his valuable recommendations.

I would like to acknowledge the National Council of Science and Technology (CONACYT) of Mexico and the National Aeronautics and Space Administration (NASA) for funding my graduate studies, under the *Bicenterario* fellowship and the ROSES NNH11ZDA001N-ACCESS project respectively. I would like to thank the Texas Advanced Computer Center (TACC) for providing access to their supercomputers. Likewise, I'd like to thank Kristin Tolle and Prashant Dhingra from Microsoft Research for providing training and access to the Azure cloud.

I would like to acknowledge Esri for giving me the opportunity of interning at their headquarters in two occasions and the great people that I met there, especially Sean Breyer, Caitlin Scopel, Daniel Siegel, Richard Nauman, Michael Dangermond, Dean Djokic, Nawajish Noman, Liz Graham, Steve Kopp, Witold Fraczek, Charlie Frye, Mark Smithgall, Michael Adams, Jessica Sweet, Andrew Norris, and Jessica Acosta-Rodriguez.

I would also like to thank my friends and colleagues Georges Comair for helping me during my first steps as a researcher and Irene Garcia-Marti for introducing me to the fields of geographic web applications and web development. I would like to thank my friends Jorge L. Loor and Gabriel A. Parra with whom I developed a special bond and

friendship over my course of time in Austin, and I would like to thank Jennifer Paz for her encouragement during the writing process of this dissertation.

Lastly, I would like to thank all students at the Center for Research in Water Resources (CRWR), especially to Xing Zheng, Cyndi Castro, Cassandra Fagan, Marcelo Somos, Prabhas Gupta, Fernando Salas, Harish Sangireddy, Allison Wood, Elisabeta Poci, Samuel Sandoval-Solis, Arthur Ryzak, Alfredo Hijar, Denny Rivas, and Carlos Galdeano.

Large-Scale Statistical Analysis of NLDAS Variables and Hydrologic Web Applications

Gonzalo Enrique Espinoza Dávalos, PhD

The University of Texas at Austin, 2016

Supervisor: David R. Maidment

The Land Data Assimilation System (LDAS) is a model developed by the National Aeronautics and Space Administration (NASA) for the purpose of quantifying the heat and water fluxes between the atmosphere and the land-surface hydrology. LDAS has two forms: National (NLDAS) and Global (GLDAS). The NLDAS grid is 1/8° with hourly and monthly estimates since 1979. The LDAS model output provides a comprehensive time-space dataset. A statistical analysis is necessary to obtain descriptive information, understand seasonal patterns, spatial distribution, and frequency distribution of the model output. The current conditions can be compared to those in the past by using statistical distributions for each variable unique to each time interval and spatial grid point. This dissertation objectives are: (1) perform a statistical analysis on the time series of NLDAS variables and model their spatial-temporal probability distributions, (2) improve data exposure through the comparison of current values with the past using web applications, and (3) evaluate the framework for access to NLDAS data. The methodology presented consists of: (1) the estimation of the NLDAS cumulative distribution functions (CDFs) on a daily and a monthly time step and development of the probability models for five variables: precipitation, runoff, soil moisture, evapotranspiration, and temperature. (2) The creation of dynamic websites displaying the

maps, time series, and latest values in the NLDAS model and its relation with the historic distributions. And (3) the implementation of time-indexed and spaced-index data access procedures. The methodology is implemented using the latest technologies in high-performance computing (HPC), cloud storage and deployment, and Geographic Information Systems (GIS) that allow performing this analysis on a large dataset (NLDAS) on a national scale, using the United States as a study case. A statistical analysis of the NLDAS model output and the comparison of current values with the historic distribution provides a thorough insight of the ranges, extremes, and seasonal variation of the hydrologic variables. The exposure of large scientific datasets such as NLDAS though the use of standards and web applications can enhance its use in hydrologic sciences and engineering.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 SPACE, TIME, AND UNCERTAINTY

Hydrologic data provide a measurement or estimate of the value of a variable in the hydrologic cycle. Hydrologic variables describe spatial-temporal processes, the result of an intricate interaction from large scale climate dynamics to local conditions. This complex system involves a large degree of uncertainty, so hydrologic variables are considered stochastic: variables with underlying distributions or parameters that vary in time and space.

An understanding of the spatial-temporal nature of the hydrologic processes, improves the assessment of natural hazards (e.g. flood or droughts) which are extreme events in a hydrologic probability distribution. Furthermore, hydrologic variables are intrinsically interrelated: one hydrologic process is preceded and followed by another. For instance, rainfall precedes surface runoff which leads to river flow. Moisture in the soil is preceded also by rainfall and followed by evapotranspiration or infiltration. This implies that each hydrologic variable has essential information about other related variables, and in combination they can be used for simulation of a state in the hydrologic cycle.

Each hydrologic variable is a random field in the space-time continuum, interrelated with the random fields of other variables. This fields and their interrelation are mathematical descriptions of the interactions between the land and the atmosphere. A thorough examination of the spatial-temporal variability of hydrologic variables should include the estimation of uncertainty through statistical analysis of the probability distributions.

1.2 HYDROLOGIC DATA COLLECTION

Traditionally, hydrologic data has been collected at discrete locations (stations). Measurements are usually taken at regular time intervals, although gaps of data occur due to malfunctions or equipment maintenance. In-situ measurements provide data for an extended period of time, but it is common to have a low station density (i.e. a few stations in a large area). Sparse observational data are available in remote areas or in developing countries.

In addition, hydrologic measurements can be made using indirect methods. For example, estimating rainfall from radar or estimating water flux from satellite records of the gravity field. In general, these datasets cover large extents, covering continents or the whole world and are continuous in space. The limitations are reduced spatial resolution and short horizons of time. Furthermore, these indirect methods have to be calibrated to specify the relationship between the variable measured and the variable estimated or modeled.

Other types of hydrologic datasets are the outputs of models. Hydrologic models are quantitative representations of the water cycle, in which assumptions and simplifications have been made in order to express mathematically the interrelationships among complex hydrologic processes. The advantages of using models are: (1) avoiding incomplete datasets, (2) providing data where it is not observed, and (3) having more consistency in the data. In contrast, a disadvantage of using models is to add errors to our estimations. A desired model would have small or negligible errors. The reduction of the estimation error is made though calibration of the model at points with known values or in-situ measurements.

1.3 STATISTICAL ANALYSIS OF LAND-SURFACE MODELS

Land-surface models provide estimates of hydrologic variables for a given point in space and period of time. These estimates are computed from forcing atmospheric parameters and mathematical representations of the fluxes between points in space and time. These deterministic values are relevant in their own right but additional information can be derived from them. Similar places in time and in space (i.e. same geographic location and same time of the year) might have different estimates; the compilation of these values over time creates a large sample which can be analyzed statistically, to describe the common values, their distribution, range, extremes, and variability between other factors.

The statistical analysis of land-surface model creates an added value for the product. Each point in time and space has an estimated value but also a probability distribution, derived from the historic values. These probability distributions, which are a function of the past, are associated with current values in order to compare how likely or extreme their values are. The result of the statistical analysis can also be used in combination with new or forecasted values, to detect large anomalies or provide quality control of the forecasted data.

1.4 WEB SERVICES AND APPLICATIONS

The rapid development of information technologies has a deep impact on the distribution and presentation of hydrologic data. Large-complex datasets can be shared in interactive maps and web applications. The advantages of using web applications are (1) the synthesis of information in a single website, (2) the immediate access to latest conditions, and (3) the quick identification of risk areas. The downside of using web applications is the high-level of expertise required to develop them and the complications of setting up the interaction between different web services and technologies.

The display of hydrologic data through dynamic and interactive web applications instead of static results images has an intrinsic added value: the user has more control over the information displayed and it can be used in 'what-if' scenarios for natural disaster assessment. The development of web-applications integrating web services and geographic data is a complex coordinated process. All the parts must be perfectly synchronized to create a product that is useful, valuable, simple, and accessible.

1.5 RESEARCH QUESTIONS

Further analysis can be made in a land-surface model output. Specifically, statistical distributions could be computed and correlated within variables providing more information about the variation of the distribution with time and space. Using this schema, a better estimation for assessment of extreme events can be made. The results of the statistical analysis can be shared though web applications, summarizing the current conditions within historic context.

In order to perform an exhaustive statistical analysis in an extensive dataset such as NDLAS, the information technologies for querying, accessing, and retrieving the data must be optimal. In this case, it is feasible to replicate the analysis in an efficient way. An optimal system will include the use of spatial maps, time series at a points (also known as "data rods"), and standardized web services.

The statistical analysis of the historic time series and the exposure of the results in dynamic websites can leverage our understanding of extreme events which would be valuable base information for future projects studying floods or droughts.

In consideration of the previously stated areas of opportunity, three main research questions are raised:

1. *How can a multi-dimensional analysis of land-surface models improve our understanding of the distribution of hydrologic variables?*

   A spatial-temporal statistical approach is performed to analyze and interpret the outputs from a land surface hydrologic model (North-American Land Assimilation System NLDAS).

2. *How can hydrologic information be shared dynamically as a final result in an accessible, simple, and interactive approach?*

   A web application architecture is constructed using the best and latest technologies available; linking web services, cloud deployment and storage, and mapping.

3. *How can large models datasets be queried, parsed, and used efficiently in hydrologic analysis?*

   A detailed web-based process for data retrieval and integration is described and implemented.

The first line of research analyzes thoroughly NLDAS model's output through the estimation of statistical parameters (such as mean, variance, and percentiles) for each point in space and time, and fitting probability distributions to its variables.

The second objective is focused on practical applications for sharing spatial-temporal data. The implementation includes the latest technologies in web applications, web GIS, cloud storage, and web services. The web applications developed are successful case studies in the highly evolving field of web GIS.

The third line of research establishes the foundation in which large datasets can be shared, queried, parsed, and used in research. The objective is to improve data access performance through the use of web services and standards, leveraging its use and dissemination in applied engineering and research.

1.6 RESEARCH OBJECTIVES

The scope and objectives of each research question are the following:

1. Complete a statistical analysis of the NLDAS model output. The statistical analysis includes the summary of the statistics and the calculation of the cumulative distribution functions (CDFs). This is performed for each grid point and for each calendar day and each calendar month, and modeling the CDFs is accomplished using common probability distribution functions. The statistical analysis covers the continental United States, using data from 1979 to 2013 and five variables (soil moisture, evapotranspiration, precipitation, runoff, and temperature) on a 1/8 degree grid with one-day and one-month time steps.
2. Create three web map applications for exposing the latest results in NLDAS: (1) latest conditions in soil moisture in Texas and its comparison with the historic trend, (2) statistical map for the continental United States showing the latest conditions and its comparison with historical values for five hydrologic variables (i.e. soil moisture, evapotranspiration, precipitation, runoff, and temperature), and (3) a time series (i.e. data rods) explorer for improving data access and displaying of LDAS data (NLDAS and GLDAS) and two additional global datasets: the Tropical Rainfall Measuring Mission (TRMM) and the Gravity Recovery and Climate Experiment (GRACE).
3. Describe a detailed framework for accessing NLDAS data. This focuses on improving performance depending on the application case using two alternatives: for space-indexed or time-indexed data. Two study cases are carried out: (1) comparing current conditions with long-term historic

trend, where space is the main variable and (2) the use of “data rods” (i.e. time series constructed from a given point in space) as input in hydrologic routing, where time is the main variable.

Chapter 2: Literature Review

The field of statistics is a core part of hydrology due the uncertain nature of climate and hydrologic variables (Maidment, 1993). Traditionally, frequentist statistical approaches have been applied to hydrologic records obtained from in-situ station, such as the mean stream discharge in a river or the normal precipitation for a given month of the year. A multidimensional (two dimensional in space i.e. latitude and longitude plus a time dimension) statistical analysis can be performed on a land-surface model, computing probability distributions for hydrologic variables that are functions in a space-time continuum. These probability distributions describe random fields of hydrologic variables (Vanmarcke, 2010).

The accelerated and ever-expanding advance in information technologies allows us to formulate novel approaches to the dissemination and analysis of hydrologic data. These new approaches can be implemented on large datasets using High-Performance Computing (HPC) and novel technologies that allow instantaneous access to data through the web. Data can be queried and displayed automatically, enabling real-time description of climate events and producing valuable information during natural disasters (Rajkumar, Lee, Sha, & Stankovic, 2010).

In summary, the present research focuses on (1) performing an statistical analysis on large-scale hydrologic datasets, (2) exposing the results on dynamic web applications, and (3) improving access to hydrologic data through web-based systems and its usability in applied research,. The data used in this research is from the North American Land Data Assimilation System (NLDAS) for the continental United States and for a period of time of 35 years (from 1979 to 2013).

2.1 THE NORTH-AMERICAN LAND DATA ASSIMILATION SYSTEM (NLDAS)

The Land Data Assimilation System (LDAS) is a compilation of Land-Surface Models (LSM) datasets that provide data for hydrologic variables in a time-space continuum with regular intervals (Mitchell, 2004; Xia, Ek, Wei, & Meng, 2012). The strength of LDAS relies on the forcing parameters (derived from observations) that reduce the bias and error of land-surface models and are derived from atmospheric models. LDAS is an extensive validated hydrologic dataset (Xia, Mitchell, Ek, Cosgrove, et al., 2012; Xia, Mitchell, Ek, Sheffield, et al., 2012). It provides continuous information which can be used as a deterministic model but it also can be used to identify trends, statistical distributions, and the strength of the relationships between variables (Rodell, Mocko, & Beaudoing, 2015). The LDAS dataset is extensive but it can be queried and accessed online with standard filters: location, time and variable of interest in an automatic way Rui et al., (2011).

There are two types of LDAS models: the North-American LDAS (NLDAS) and the Global LDAS (GLDAS) Rodell et al., (2015). The NLDAS spatial coverage includes the continental United States and buffer areas into the northern and southern borders. It has a spatial resolution of 1/8°. NLDAS temporal coverage is from 1979 to the present and in an hourly resolution. In contrast, The GLDAS model spatial coverage is the world but in a coarser resolution in time (3 hours) and space (1/4° grid). (Goddard Earth Sciences Data and Information Services Center, 2015; M Rodell et al., 2004)

NLDAS includes four main different LSMs: (1) Noah (Acronym formed by the four institutions that develop it: *N*ational Centers for Environmental Prediction, *O*regon State University, *A*ir Force, and *H*ydrology Lab from the National Weather Service), (2) VIC (Variable Infiltration Capacity), (3) Mosaic, and (4) CLM. The differences between

the models relates to the assumptions or simplifications in the moisture and energy fluxes at the land surface (Mitchell, 2004).

From the models in NLDAS, the Noah model is selected for the present research because of its extended use in hydrologic sciences and the accessibility of the data through web services. The Noah LSM was derived from a less complex model developed at Oregon State University in the 1980s which has been subject to revisions and improvements since then Ek et al., 2003. The NLDAS-Noah dataset is indexed by space in the Grid Application Development Software (GrADS) (Mitchell, 2004) and as a Web Map Service (WMS) through the Giovanni Portal (Rui et al., 2011, 2013). This means that a file with the spatial coverage over the whole domain can be obtained for a single time-step (Berman et al., 2001). The WMS is a standard of the Open Geospatial Consortium (OGC) (de la Beaujardiere, 2006) which facilities its use across platforms in a systematic and automatic way. Similarly to the spatial-indexed data, the NLDAS-Noah data is also available in a web service indexed by time called "data rods" (Rui et al., 2013), which means that the time-series for a given point in space can be obtained in standard formats (WaterML, NetCDF, or NetCDF). The double indexation by time and space optimizes the process of accessing the data.

One of the reasons of choosing NLDAS is the advantages that LSM have in hydrology over Global Circulation Models (GCMs). The study made by Jiang, Gautam, Zhu, & Yu, (2013) shows that GCMs have problems replicating extreme precipitation values, especially high precipitation values in short periods of time. The research states that the use of these models is improper for flood and drought assessments and suggests that LSM, uncoupled from atmospheric models, and calibrated with in-situ measurements could improve the results.

Lakshmi, (2004) identifies the challenges in hydrologic sciences for the estimation of variables in ungauged basins. He acknowledges the importance of satellite products and land-surface models for improving estimations, and emphasizes the need for studying and mapping soil moisture for accurate water balances (Espinoza, Arctur, Teng, et al., 2015). The study performed by Lakshmi, Piechota, Narayan, & Tang, (2004) used soil moisture data from the VIC model as an indicator of hydrologic extremes (i.e. floods and droughts) in the upper Mississippi basin. The research shows that the analysis of long term soil moisture and its anomalies can be used as a drought indicator. They identified mean and common ranges of soil moisture values during normal, flood, or drought conditions but did not associate probabilities for these values (Espinoza, Arctur, Teng, et al., 2015). A statistical analysis of the LSM data is needed to associate probabilities with modelled values, leveraging the understanding of extreme events.

An vast precipitation dataset is provided by the Global Precipitation Climatology Centre, (2016) that is part of the World Meteorological Organization (WMO) and managed by the German Meteorological Service (DWD). The dataset is an extensive compilation and interpolation of precipitation records across the world, it has a monthly temporal resolution and different products with varying spatial resolution (e.g. 0.5°, 1.0°, 2.5°) and temporal coverage (e.g. since 1951, 1982, 2004).

2.2 STATISTICAL ANALYSIS OF HYDROLOGIC VARIABLES

The analyses of the spatial-temporal statistical dispersion of hydrologic variables, their common range, and their expected values are part of the major applications of statistics in hydrology. The research of El Adlouni, Bobée, & Ouarda, (2008) examines the common distributions used in hydrology and classifies them accordingly to their tail behavior. The research reinforces the importance of the tail behavior for the estimation of

extreme events. Therefore, they conclude that fits that performs well at both ends of a CDF is highly desirable. The study performed by Coles, Pericchi, & Sisson, (2003) uses rainfall data from coastal and central Venezuela and proves that extreme rainfall events though exceptional can be properly estimated, if the rainfall probability distributions are constructed.

The probability distribution of hydrologic variables can be estimated using the complete time series or partial, minimum, or maximum series. Beguería, (2005) focuses on modeling extreme rainfall events using the series of annual maximum and partial duration. This method requires defining threshold values which can be complicated and subjective to estimate. The percentiles of the distributions are affected by these thresholds (Katz, Parlange, and Naveau 2002), that can lead to subjective results. If sufficient data is available, the complete time series can be modeled using a probability distribution, reviewing the performance of the model at all the range of values, while attending special consideration at the extremes.

Entekhabi & Rodriguez-Iturbe, (1994) studied the hydrologic processes from a water balance perspective. The study emphasizes in the spatial-temporal scales of the different hydrologic processes, focusing in infiltration, soil moisture, precipitation, runoff, and evapotranspiration relating changing the variability at different scales using filters in the time series with. It shows that computing averages over time and space can affect the statistical fields for rainfall intensity and soil moisture, a statistical analysis over the complete time series is more desirable. Viglione et al., (2010) studied the impact spatial-temporal precipitation and runoff variability for historic flood events in northern Austria. The study concludes that the spatial-temporal characterization is necessary to understand the hydrologic system and that additional variables, such as soil moisture are relevant at larger scales.

New, Hulme, & Jones, (1999, 2000) created a dataset showing the monthly values of hydrologic variables (such as precipitation, mean temperature, vapor pressure, etc.) in a half-degree cell size grid for the world. These studies used mean and standard deviation values per cell and time to compute anomalies but did not associate them with a probability distribution. The dataset was created interpolating observation data that induced an additional error. The use of NLDAS data could improve the spatial-temporal mean and standard deviation in a finer grid (1/8°) and the calculation of the probability distributions creates an added value to the dataset.

The study performed by Husak, Michaelsen, & Funk, (2007) is similar the present research. It uses rainfall data from the Collaborative Historical African Rainfall model (CHARM) (Funk et al., 2003) and fits probability distributions to each grid cell to improve monthly estimates. The study shows that the gamma distribution properly represents the historic values on more than 98% of the sites and improves the estimation of the cumulative distribution functions. The CHARM data is coarser than NLDAS with half-degree cell size, the research downscaled the data (to a 0.5° grid) using an underlying Digital Elevation Model (DEM) which can induce additional errors. The present research takes the key ideas from Husak et al., (2007) such as fitting distribution per cell and per time step, but the statistical analysis is expanded allowing fitting distributions on a smaller time step (daily), and in some cases with a two-step process (Section 3.2.3).

Groisman et al., (1999) modeled daily and monthly precipitation data with a gamma distribution for selected countries including the United States. The study concluded that (1) the gamma distribution is appropriate for modeling daily precipitation data, (2) the parameters vary greatly with time and space, and (3) an increase in mean precipitation can lead to more frequent heavy rain events. From the research, is inferred

that the computation of the spatial-temporal mean precipitation values is needed as baseline data for studying increase in frequency of heavy storm events. In addition, the study of Wigley, (2009) estimates how a change in the mean of a distribution can affect the values of an extreme event. It shows how an increasing mean can reduce the return period of an event and that the change in mean represents a challenge for frequentist statistics.

Wang, McKenney, Shang, & Li, (2014) studied the differences in water budget imbalances using long-term gridded precipitation and evapotranspiration data, using Canada as a study area. The imbalances were greater in regions with sparse stations, in which the use of land-surface models could be beneficial. The research did not take a statistical approach for the computations of the water balances which could provide a range of likely values and improve the estimator (Coles et al., 2003).

The goodness-of-a-fit test used in Husak et al., 2007 was the Kolmogorov-Smirnov (KS) test described by Darling, (1957). The KS test is based on the maximum difference between a theoretical CDF and an empirical CDF, comparing the empirical CDF or a sample from it comes from a theoretical distribution The KS test is specially suitable for the present research because it checks the fit in the entire distribution (including extremes) and not only if the mean or variance are from the same theoretical distribution (Sager, 2010). One of the limitations is the relatively complex estimation of the p-value due the distribution of the KS statistic. Marsaglia, Tsang, & Wang, (2003) developed an algorithm to approximate the p-value with at least seven-digit accuracy, which have been implemented in the R language (R Core Team, 2014).

The website developed by the Australian Goverment - Bureau of Meteorology, (2016) is similar to the present research. The website shows the daily values for different hydrologic variables such as soil moisture (at different levels), actual and potential

evapotranspiration, precipitation, and runoff. The results are presented in an interactive web application that shows the actual or relative conditions. The data comes from the Australian Water resources Assessment Landscape (AWRA-L v5.0) which is a LSM with a 0.05° resolution (Vaze et al., 2013).

2.3 GEOGRAPHIC INFORMATION SYSTEMS (GIS), INFORMATION TECHNOLOGIES AND DATA ACCESS

The paper presented by Beniston et al., (2012) is a thorough description of the data access challenges faced in water resources research. In summary, the challenges are (1) the frequent time-space scarce datasets, (2) gaps in data, (3) availability, (4) charges involved on sharing datasets, (5) disparity between data availability in successful socio-economical areas and areas with limited resources, and (6) the use of standards in storing and sharing information. The use of NLDAS in the present research minimized all the challenges due (1) the dataset is continuous in time and space. (2) There are no gaps of data in the model. (3) The data is available through web services. (4) The dataset is public, and (5) the cell size and time interval are the same for all estimations in the dataset. The present research focuses on improving (6) exposing the results of the statistical analysis through web applications, standards, and cloud storage.

The study performed by Dragićević, (2004) is a recapitulation of web-based Geographic Information Systems (GIS). The research identifies the three main areas of new developments in web-based GIS:

> *"(1) Spatial data access and dissemination, (2) spatial data exploration and geovisualization, and (3) spatial data processing, analysis and modeling".*

The present research is based on these three main areas and their integration in simple and useful web applications.

The use of geospatial standards is essential for sharing data online. Standards that are developed by the Open Geospatial Consortium (OGC) are largely accepted by the geospatial scientific community. Primarily, because OGC is an organization formed by commercial companies, government institutions, and universities that seeks to develop and approve information technologies that leverage geospatial research, Open Geospatial Consortium, (2014). OGC approved standards that are cross-platforms, such as Geographic Information Web Services for online mapping (Alameh, 2003) or WaterML for distribution of hydrologic time series data (Valentine, Taylor, & Zaslavsky, 2012). The use of these standards improve the process of sharing information and set the rules of accessing data online (Stollberg & Zipf, 2007). A key advantage of standardization is the ability to automate data access.

An example of online sharing of hydrologic information is being made by the Global Earth Observations System of Systems (GEOSS), part of the Group on Earth Observations (GEO). GEOSS is a successful example of management, sharing, and analysis of global datasets, aiming to provide scientific answers to global environmental problems in benefit of the society (Lautenbacher, 2006).

An example of the use of standards by Botts, Percivall, Reed, & Davidson, (2008) is the implementation of field sensors with trigger threshold values. When a threshold value is surpassed an alert is sent, displaying vulnerable areas in a geospatial context. A similar approach is adopted in the current research, where the developed web applications can show the latest results in NLDAS and identifying areas where the anomaly is large.

Furthermore, current research developments focus on geoprocessing services in addition to the traditional web mapping and feature web services. Geoprocessing services are capable of executing tasks on a server, returning only outputs to the client. The advantages are the centralization of the computer power needed and the ability of using

already set-up workflows (Michaelis & Ames, 2009). In the present research, the analysis is made using High-Performance Computing (HPC) resources (Texas Advanced Computing Center & The University of Texas at Austin, 2015), allowing to process 35 years of data, for 5 variables, and for the continental United States in an efficient way.

A successful example of integration of GIS, data services, HPC, and large-scale modeling was carried out by Maidment, (2015) titled the National Flood Interoperability Experiment (NFIE). The goal of the NFIE is to improve flood forecasting and emergency response at high spatial resolution for the continental United States. It combines data from the National Hydrography Dataset (NHD) and the National Weather Service (NWS) with forecast models such as the Weather-Research and Forecasting model (WRF-Hydro), Rapid, and HEC-RAS to predict stream discharge and in some areas the flood inundation map. The implementation and development of projects like the NFIE are deeply founded in the advancement of information technologies.

Lastly, the emerging field of CyberGIS (Liu, Padmanabhan, & Wang, 2015) is the evolution of GIS on the web. It studies all parts required for a spatial analysis on the web: data services, spatial services, geoprocessing services, online modeling and analysis, and infrastructure. CyberGIS focuses on interactive solutions that rely on large geospatial datasets and its integration with other networks. The developed web applications emerge as great prototypes on the statistical and hydrologic fields.

2.4 SUMMARY

2.4.1 State-of-the art

The baseline developments in which the current research is established are (1) the standardization of web services for sharing information through the internet. (2) The accessible web-based alternatives of retrieving the information as time-indexed or space-

indexed. (3) The statistical modeling of hydrologic variables, which is extensively used in the field, and the calculation of probability distributions. (4) The estimation and fitting of statistical distributions to continental-scale data, allowing spatial-temporal variation. (5) The growing use of Global and National-scale Land-Surface Models (LSM) as base data. And (6) Web GIS, on-the cloud storage, and interactive web applications are the new developments in data technologies, which can be used to connect web services, geographic components, and data under a common framework.

2.4.2 Gaps in knowledge

The existing gaps in knowledge are the following:

- A thorough statistical analysis of Land-Surface Models (LSM), including the modeling and interpretation of the underlying spatial-temporal statistical distributions at a national scale.
- The improvement of hydrology data exposure in web applications. Migrating from websites where the information is deep and hardy accessible, to a more direct way of sharing information through informative websites. Displaying latest results and is comparison with historic distributions.
- An extensive description and establishment of procedures for data retrieval and data querying. Evaluation of the performance and direct integration within software and analysis in research.

2.4.3 Scope and contributions of the research

The contributions of this research are (2) the development of a methodology to analyze the statistical distributions of hydrologic variables in a LSM across time and space, and to compare the latest results available with them. (2) The creation of

hydrologic web applications for exposing data online, using leading-edge technologies. And (3) the description, evaluation, and integration of web-based frameworks for accessing LSM data.

The research scope includes the statistical analysis and the fitting of probability distributions in a day and month basis for five NLDAS variables: soil moisture, evapotranspiration, precipitation, runoff, and temperature. It also includes the development and deployment of web applications that serve as integration of web mapping and data services, and a thorough description an evaluation of the data access framework of NLDAS data using time-indexed and space-indexed servers.

The scope of the research does not include the study of the effects caused by climate change on the historic distributions; it does not consider the downscaling of NLDAS data in a finer spatial resolution; and the web applications shows the latest results in NLDAS and not forecast values.

Chapter 3: Multidimensional Statistical Analysis of Hydrologic Parameters[1]

The objective of the methodology of this chapter was to model the empiric probability distributions of hydrologic variables through common family distributions. The modeling of the probability distributions was made through the calculation of the empirical Cumulative Distribution Functions (CDFs) on a daily and a monthly basis per variable. Five hydrologic variables were considered: (1) soil moisture in the top meter (kg/m^2), (2) total evapotranspiration (kg/m^2), (3) surface runoff (kg/m^2), (4) precipitation (kg/m^2), and (5) two meters above ground temperature (K). Furthermore, the empirical CDFs were fitted by a common probability distribution at each NLDAS point and each calendar day and month. Finally, the fit was validated and the result of the analysis was stored in a cloud storage service.

The resulting CDFs distributions are functions of the variable, geographic location, and time of the year. These distributions were validated and depend only in two parameters: the mean and the variance; the original data was significantly reduced into a simplified set of parameters. The web applications are described in Chapter 4 access and display the data from the cloud storage service.

3.1 METHODOLOGY

Figure 1 shows the methodology for the statistical analysis which can be divided in the following steps:

- Data acquisition (Chapter 5)
- Empirical CDFs calculation

[1] Portions of this chapter and its corresponding literature review on Chapter 2 were first published in: Espinoza, G., Arctur, D., Teng, W., Maidment, D., García-Martí, I., & Comair, G. (2015). Studying Soil Moisture at a National Level through Statistical Analysis of NASA NLDAS Data. Journal of Hydroinformatics. Accepted for publication on November 2, 2015.

- CDFs fitting
- Validation of the fits
- Results in-cloud storage

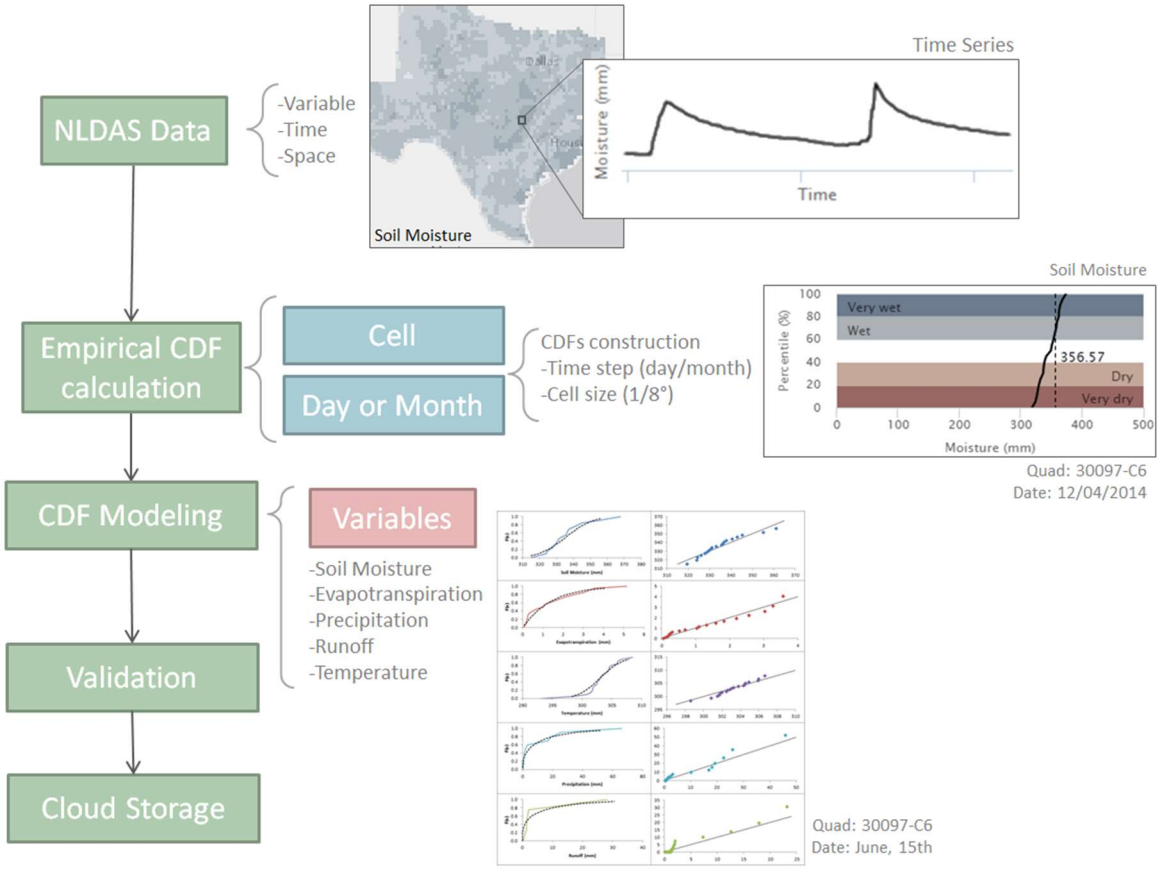


Figure 1: Description of the methodology for the statistical analysis.

3.2 Empirical CDFs calculation

The empirical CDFs were computed for each of the five variables (i.e. soil moisture, temperature, evapotranspiration, precipitation, and runoff) and for each calendar day or each calendar month. The data was obtained using the data rods service. The CDF calculation was different depending on the type of the variables: quantities (soil

moisture and temperature) or fluxes (evapotranspiration, precipitation, and runoff), and the time interval: daily or monthly.

Table 1 summarizes the different CDFs calculation types. For the variables that represent quantities (i.e. soil moisture and temperature), the CDFs were computed for the mean values for a given day or month. On the contrary, for the variables that represents fluxes (i.e. evapotranspiration, precipitation, and runoff), the CDFs were computed for the cumulative values for a given day or month.

Hydrologic variable	Day interval		Month interval	
	Values type	Calculation	Function type	Calculation
Soil moisture	Mean	One-step	Mean	One-step
Temperature	-	-	Mean	One-step
Evapotranspiration	-	-	Cumulative	One-step
Precipitation	Cumulative	Two-steps	Cumulative	One-step
Runoff	Cumulative	Two-steps	-	-

Table 1: CDFs calculation types for each variable based on the type: flux or quantity, and the time interval: day or month.

For the time interval, the CDFs calculations were one-step (Section 3.2.2) or two-steps (Section 3.2.3). The one-step calculation was made directly from the time series. In contrast, the calculation in the two-steps process removed the large amount of zeros in the time series (i.e. for precipitation and runoff variables in the day interval) before computing the CDFs. First, the event of precipitation/no precipitation (or runoff/no runoff) was treated as a Bernoulli trial. The second step was the CDF calculation of

precipitation or runoff given that the event occurred (success in the Bernoulli trial), on the time series without the zero values.

3.2.1 Mathematical definitions

The definitions of the mathematical terms in the CDFs calculation are the following:

- $x(t) = x$: Value of the hydrologic variable x (e.g. evapotranspiration) at a time t.
- t: time stamp of an hydrologic measurement, it can be a month or a day depending on the time interval (e.g. August-2010 or 8/12/2010)
- $X = X(t)$: Value of the hydrologic variable x at a time t.
- d: Calendar day or calendar month (e.g. June 8th or February).
- $S_x|_{t_1}^{t_2}$: Set of values (time series) for the values of x in the time interval $[t_1.t_2]$.
- $S_{x,d}\big|_{t_1}^{t_2}$: Subset of values from the time series $S_x|_{t_1}^{t_2}$ that match the calendar day or month d.
- n: Number of elements (days or months) in $S_{x,d}\big|_{t_1}^{t_2}$.
- $n_d|_{t_1}^{t_2}$: Number of days with a precipitation or runoff event in the time interval $[t_1.t_2]$. [2]
- evt_d: denotes if a precipitation or runoff event occurred in a calendar day d, it only takes *True* or *False* values (Boolean variable).[2]
- $P_{evt}(d)$: Probability of a precipitation/runoff event at the calendar day d.[2]
- $O_{x,d}\big|_{t_1}^{t_2}$: Ordered set of $S_{x,d}\big|_{t_1}^{t_2}$. Where:
  - The first element $O_{x(1),d}\big|_{t_1}^{t_2} = \min\left(S_{x,d}\big|_{t_1}^{t_2}\right)$

[2] Only used for the daily time interval and the variables of precipitation or runoff.

- The last element $O_{x(n),d}\big|_{t_1}^{t_2} = \max\left(S_{x,d}\big|_{t_1}^{t_2}\right)$
- And $O_{x(i),d}\big|_{t_1}^{t_2} < O_{x(i+1),d}\big|_{t_1}^{t_2}$ for $i \in [1, n-1]$

- $f_d(x) = P_d(X = x)$: Probability density of x at a calendar day or month d. This is the Probability Distribution Function (PDF) of x for a given calendar day or month d.
- $F_d(x) = P_d(X \leq x)$: Probability of X being less or equal to x at a calendar day or month d. This is the Cumulative Distribution Function (CDF) of x for a calendar day or month d.
- $q_p\big(F_d(x)\big) = q_p$: Percentile value of the variable x for a cumulative probability of p at a given calendar day or month d.

The goal of the empirical CDF calculation (one-step or two-steps) was to calculate the CDF curves for each calendar day or month d, that is obtaining the percentile values $q_p\big(F_d(x)\big)$ for all grid points in NLDAS and all calendar days and months.

3.2.2 One-step calculation

First, a subset of values $S_{x,d}\big|_{t_1}^{t_2}$ was obtained from the time series $S_x\big|_{t_1}^{t_2}$ for the calendar day or month d (Equation 1). The procedure was repeated for the 365 days or the 12 months. Second, an equal probability was set for each element in $S_{x,d}\big|_{t_1}^{t_2}$ (Equation 2) and the values were ordered (Equation 3). The cumulative probabilities $F_d(x)$ were calculated (Equation 4) for each x value. Third, the percentile values $q_p\big(F_d(x)\big)$ were calculated (Equation 5) for the cumulative probabilities from 0 to 100 every 0.05 steps, q_0 is the minimum value, q_1 is the maximum, and in most of the cases the q_i percentiles (where $i \in [0.05, 0.95]$) are linearly interpolated using the closest upper and lower values.

$$S_{x,d}\big|_{t_1}^{t_2} = \{x(t): t \in [t_1, t_2]\ and\ \text{day}(t) = d\} \text{ or} \quad (1)$$

$$S_{x,d}\big|_{t_1}^{t_2} = \{x(t): t \in [t_1, t_2]\ and\ \text{month}(t) = d\}$$

Where:

$S_{x,d}\big|_{t_1}^{t_2} \subset S_x\big|_{t_1}^{t_2}$,

$t_1 = 1/2/1979$, and

$t_2 = 12/31/2013$

$$f_d(x) = P_d(X = x) = \frac{1}{n-1} \text{ and} \quad (2)$$

$$\text{Ordered}\left(S_{x,d}\big|_{t_1}^{t_2}\right) = O_{x,d}\big|_{t_1}^{t_2} \quad (3)$$

$$F_d(x) = P_d(X \le x_i) = \sum_{x(1)}^{x(i)} P_d(X = x_i) \quad (4)$$

Where:

$\{x: x(1), x(2), x(3), \dots, x(n)\} = O_{x,d}\big|_{t_1}^{t_2}$

$$\boldsymbol{q_p} = \left(\frac{\boldsymbol{q_{p(U)} - q_{p(L)}}}{\boldsymbol{F_d(x)_U - F_d(x)_L}}\right)(\boldsymbol{F_d(x) - F_d(x)_L}) + \boldsymbol{q_{p(L)}}\ \boldsymbol{if}\ \boldsymbol{p} \in (\mathbf{0}, \mathbf{1}) \quad (5)$$

Where:

$(q_{p(U)},\ F_d(x)_U)$ is the closest upper value,

$(q_{p(L)},\ F_d(x)_L)$ is the closest lower value, and

$q_{p(L)} < q_p < q_{p(U)}$

3.2.3 Two-steps calculation

The two-step calculation was used only for the precipitation and runoff variables on the daily case, due the large number of days without a precipitation or runoff event in

the time series. The two-step calculation modifies Equation 1 for the daily case (Equation 6), where a pre-screening of the time series $S_x|_{t_1}^{t_2}$ (additional step) was made to identify the zero values and removed them from the data. The resulting subset $S_{x,d}\big|_{t_1}^{t_2}$ did not include zeros, or in practice, negligible values smaller than a threshold (e.g.$1.25x10^{-9}\ mm/day$) were removed. The probability of a $P_{evt}(d)$ (Equation 7) is defined as the probability of having a precipitation or runoff at the calendar day d.

In this particular two-steps calculation the meaning of the PDFs ($f_d(x)$) and CDFs $\left(F_d(x)\right)$ was modified (Equation 2 and 3). The distributions actually are (Equation 8 and 9) the probabilities of precipitation or runoff depth given that the precipitation or runoff event did occur.

$$S_{x,d}\big|_{t_1}^{t_2} = \{x(t): t \in [t_1, t_2]\ and\ x(t) \geq a\ and\ \text{day}(t) = d\ \} \qquad (6)$$

Where:

$S_{x,d}\big|_{t_1}^{t_2} \subset S_x|_{t_1}^{t_2}$,

$t_1 = 1/2/1979$, and

$t_2 = 12/31/2013$

$a = 0.05\ mm/day$: Threshold for precipitation/runoff events

$$P_{evt}(d) = n_d|_{t_1}^{t_2}/n \qquad (7)$$

$$f_d(x) = f_d(x|evt_d = True) \qquad (8)$$

$$F_d(x) = F_d(x|ev_{\ d} = True) \qquad (9)$$

3.2.4 Results

3.2.4.1Spatial-temporal description of the hydrologic conditions across the United States

The empirical CDFs computed are an estimate of the spatial-temporal distribution of the variables. The CDFs show the common range of values per time of the year and geographic location. Figure 2 is a map of six selected locations in which the results are shown.

Location	Latitude	Longitude
Austin, TX	30.3125	-97.8125
Portland, OR	45.6875	-122.9375
Provo, UT	40.3125	-111.5625
Redlands, CA	34.0625	-117.1875
Tuscaloosa, AL	33.1875	-87.4375
Washington, DC	38.9375	-77.1875

Figure 2: Selected locations in which the results of the statistical analysis are plotted.

Table 1 and Figure 3 show an example of the results obtained with the statistical analysis for Austin, TX on May 5. Table 1 summarizes the *mean* and *standard deviation* for the five variables, and an additional parameter *probability of event* for the variables of precipitation and runoff. Figure 3 plots the computed CDFs for the day, which provide the historic usual values and their range. The combination of these data (the statistics and the CDFs) are used to fit probability distributions (Section 3.3)

.

Variable	Mean	Standard Deviation	Probability of Event
Soil moisture (kg/m^2)	273.33	26.76	-
Evapotranspiration (kg/m^2)	2.90	0.70	-
Precipitation (kg/m^2)	6.36	7.50	0.34
Runoff (kg/m^2)	0.64	0.64	0.17
Temperature (°C)	24.35	3.67	-

Table 2: Example of the statistical parameters obtained for Austin, TX on May 5. The mean, and standard deviation are reported for all the variables, parameters that are used to fit a probability distribution. In addition, the probability of an event is reported for the variables of precipitation and runoff.

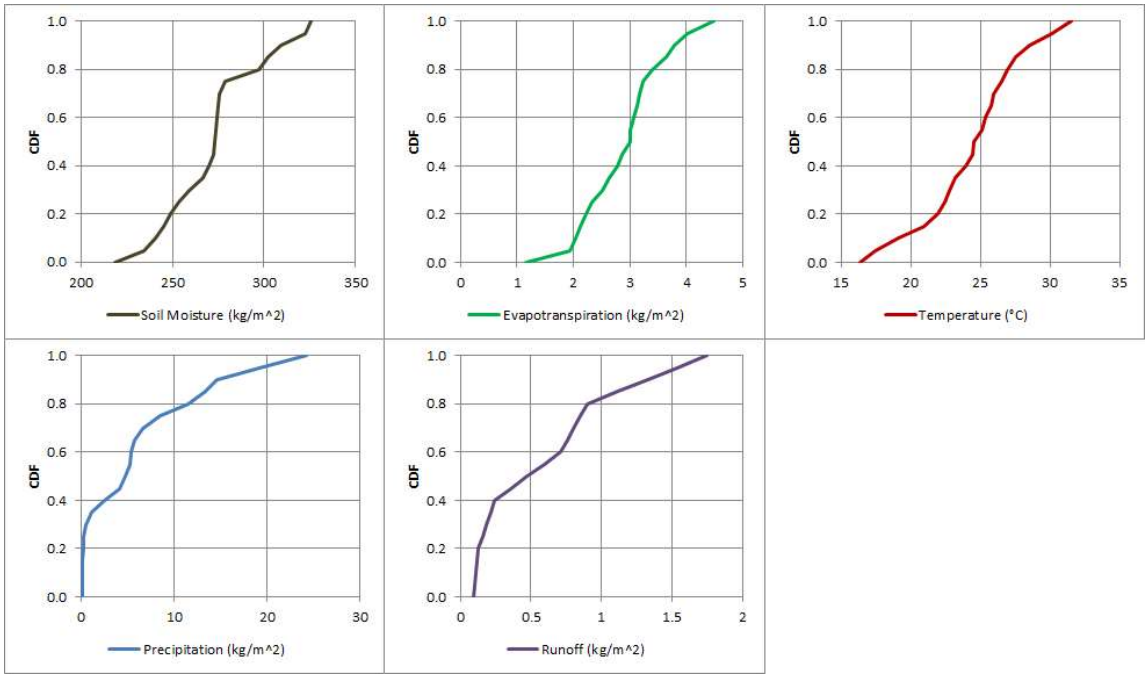


Figure 3: Example of the Cumulative Distribution Functions (CDFs) computed for Austin, TX on May 5. The CDFs are the summary of the historic conditions and they associate a probability for each value of the variables.

Figure 4 shows the CDFs for the locations in Figure 2 and for the five variables at the 15th day of each month: January, April, July, and October representing the four seasons: winter (blue), spring (green), summer (red), and fall (orange) respectively. The shape of the CDFs and the separation between different dates shows the variation and range of the variables. For example, the soil moisture distribution in Washington, DC has a narrower range of values and varies less across the year in comparison to Provo, UT.

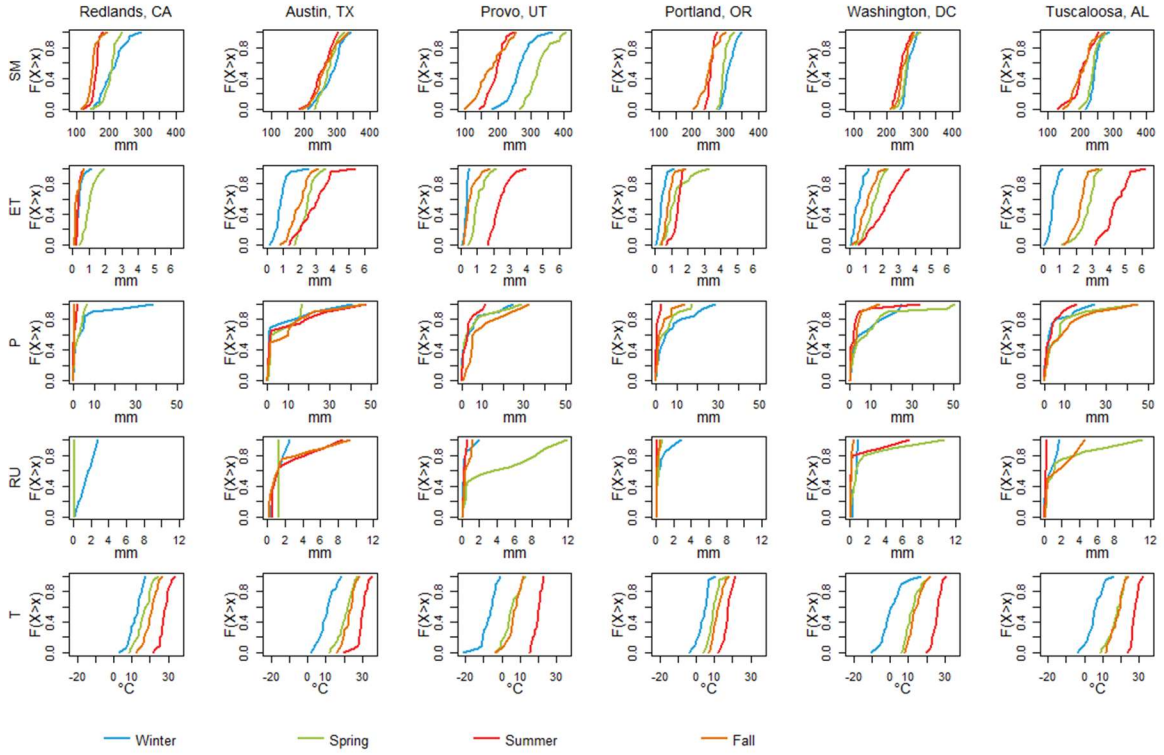


Figure 4: CDFs distributions for the variables (from top to bottom): soil moisture, evapotranspiration, precipitation, runoff, and temperature; and for (from left to right): Redlands, CA; Austin, TX; Provo, UT, Portland, OR; Washington, DC, and Tuscaloosa, AL. The CDFs represent the four season: winter (blue), spring (green), summer (red), and fall (orange), for the 15th day of the months: January, April, July, and October.

The CDFs shown in Figure 4 are indirect functions of the climate and vegetation. For example, precipitation regimes are identified in the CDFs with larger values in Austin, TX during summer and fall; in contrast Redlands, CA registers precipitation during winter. Evapotranspiration values across the year have low variability and small values in Redlands, CA. In contrast, evapotranspiration values in Tuscaloosa, AL have more variability and the range of values change significantly. Runoff increases significantly during spring in Provo, UT due melting of the snowpack.

The CDFs of evapotranspiration provide the range and expected values for each day of the year at each location, which can be useful in agriculture, water balances, and to detect when extreme values are occurring (e.g. drought conditions).

The CDFs for the variables of precipitation and runoff are clearly defined for all the locations but Redlands, CA. The large number of days without precipitation or runoff events at this location creates a smaller subset of values for the construction of the empirical CDFs. The 35-year period of data from NLDAS is not enough to construct the probability distributions at this location. This is the general case for desert and arid places in the period of time between rain seasons. In those cases, the CDFs distributions were not constructed.

Figure 5 displays the variation of the percentiles per variable across the year at each location shown on Figure 2. Lighter-green lines are for lower percentiles and darker-green for higher percentiles. The plots accurately represent the spatial variations across time. For example, the soil moisture plots capture the seasonality. Provo, UT has two distinct patterns in soil moisture characterized by a peak during spring and a valley in the fall with a constant decrement in between. Tuscaloosa, AL has almost constant soil moisture values across the year for higher percentiles but having larger variability (i.e. variance) during the summer were dryers values might occur.

Climate regimes are represented in Figure 5, the temperature plots peak during summer as expected. Evapotranspiration is a metric of the vegetation and humidity or dryness of the city. Temperature has smaller values in Redlands, CA and larger values in Austin, TX and Tuscaloosa, AL.

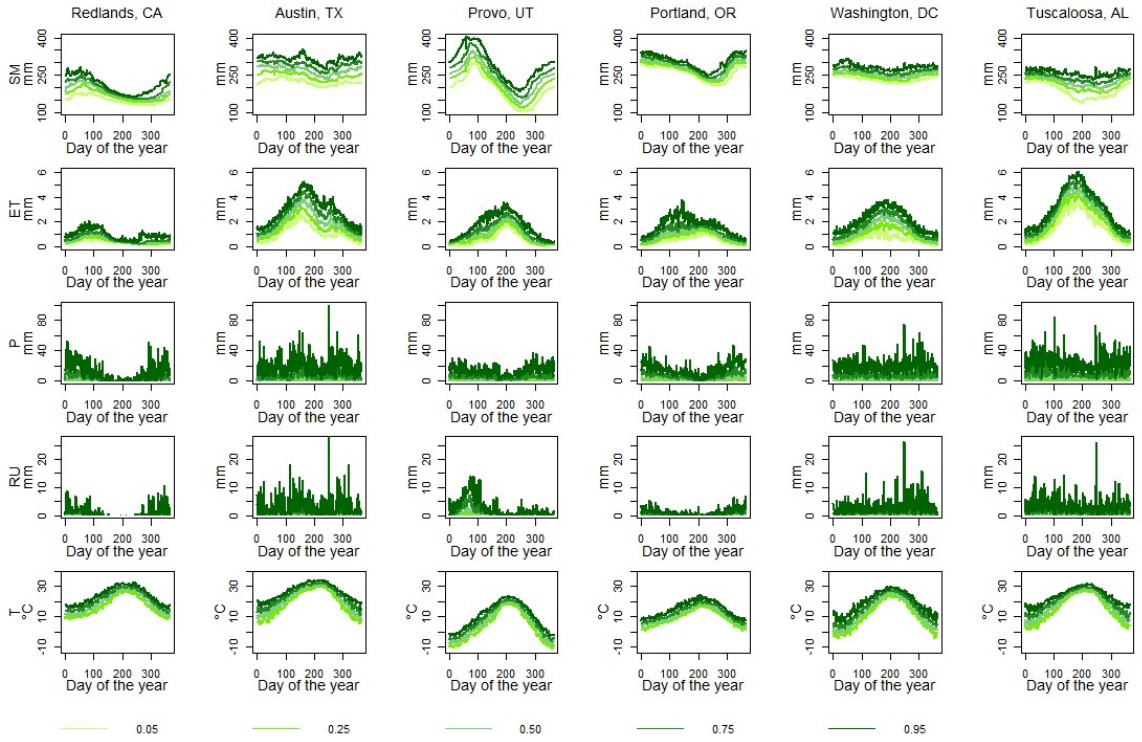


Figure 5: Variation of the CDFs across the year for five hydrologic variables (rows) at six selected locations (columns) for the percentiles: 0.05, 0.25, 0.50, 0.75, and 0.95 (from lighter to darker-green).

The precipitation and runoff plots identify periods of the year were flooding can occur and associate a probability value per precipitation or runoff depth. These values can be useful for preliminary estimates used in planning and design.

Figure 6 shows the 0.50, 0.75, and 0.95 (from light to dark purple respectively) percentile plots for precipitation and runoff through the year at the selected locations in

Figure 2. The probability of a precipitation and runoff event (dotted black line) is also shown. The statistical analysis identify the periods of time were storms are more likely to occur and when it can generate a larger runoff depth. This information can be used in addition with forecast data for flood analysis. The percentile value for the forecasted precipitation depth can be calculated, a metric of how extreme the event would be.

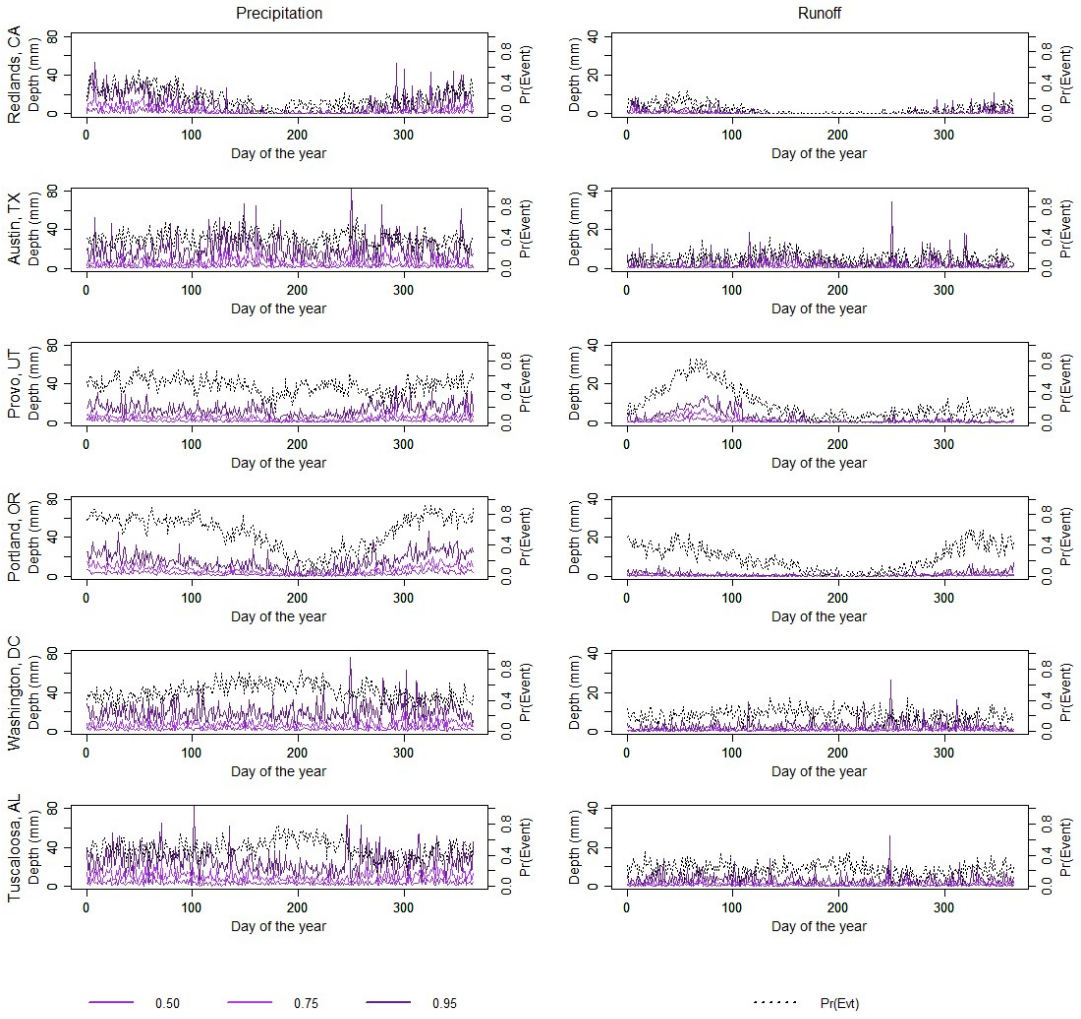


Figure 6: Precipitation and runoff depths for the 0.50, 0.75, and 0.95 percentiles (from light to dark purple) and probability of a precipitation or runoff event (dotted black line) at selected locations.

Evapotranspiration is a relevant parameter to estimate the effects of droughts in farms and vegetation. Figure 7 shows the evapotranspiration percentiles for June, 2015 in California. The Central Valley and the cost of southern California had extremely low evapotranspiration in comparison with the 35-year monthly normal (below the 0.05 percentile), which shows the severity of the ongoing drought. In contrast, the Sierra Nevada has larger evapotranspiration percentiles than in the 35-year normal (around 0.80). This increase might be due warmer conditions (Blankinship, Meadows, Lucas, & Hart, 2014).

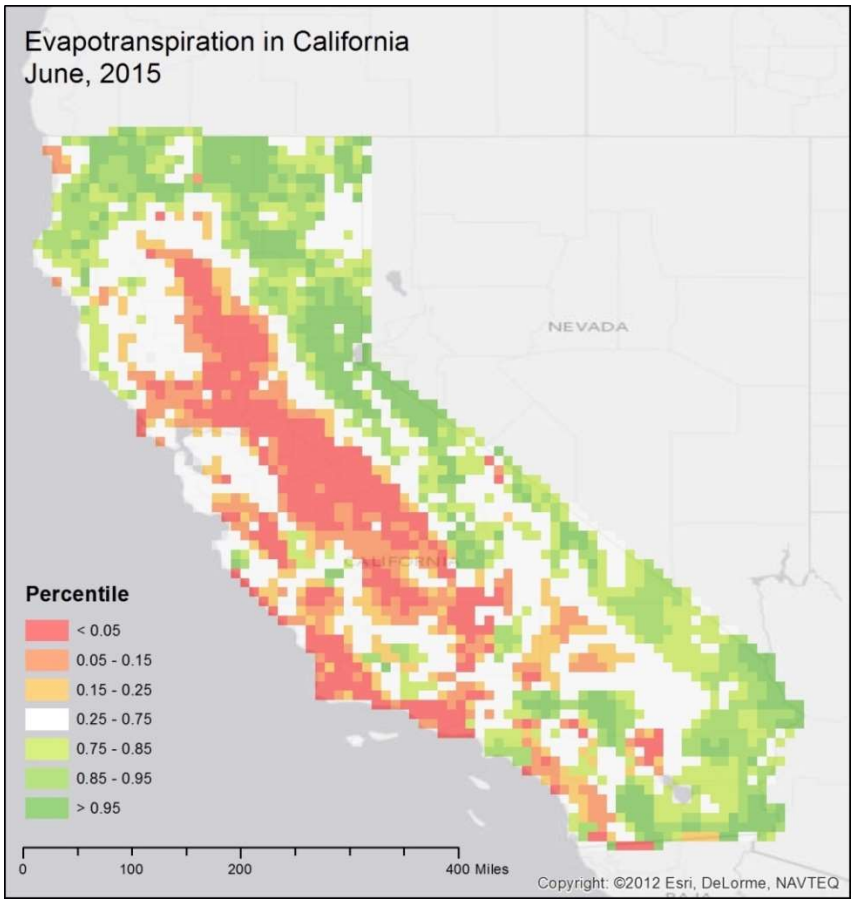


Figure 7: Evapotranspiration percentiles in California for June, 2015. The low percentile values in the central valley are a sign of the current drought conditions.

3.2.4.2 Comparison of current conditions and the CDFs for soil moisture values in Texas

The distributions are used to compare the values in the NLDAS model for a given day and the computed CDFs of that day. Figure 8 displays the CDF of soil moisture (dark line) and the actual value (dashed line) on June 13, 2014 for three locations (from top to bottom): Austin, El Paso, and Houston. The CDF for El Paso shows lower values of soil moisture due dryer conditions and a low percentile (about 18%) but it also shows a smaller range of variation and a steeper curve than for the other two cases. For Austin and Houston, the values of soil moisture are close to the median, therefore they have expected conditions.

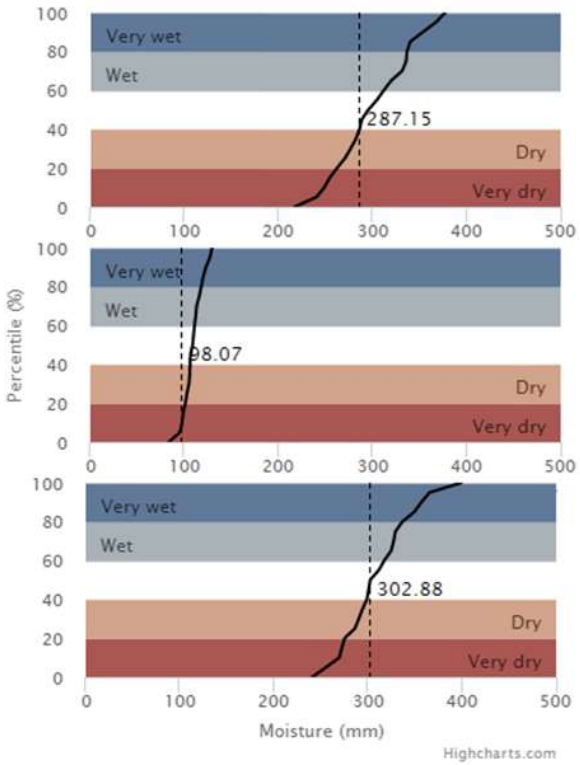


Figure 8: Soil moisture values and CDF distributions for June 13th, 2014 at (from top to bottom) Austin, El Paso, and Houston.

Figure 9 describes the monthly soil moisture CDFs for Austin. The plot show how the range of values in soil moisture varies with time. The CDFs are unfixed and time shifts the values or the shape of the CDFs. The calculation of the empirical CDFs for a given calendar day (or month) at each grid cell, properly describes quantitatively the uncertainty on a variable across time and space.

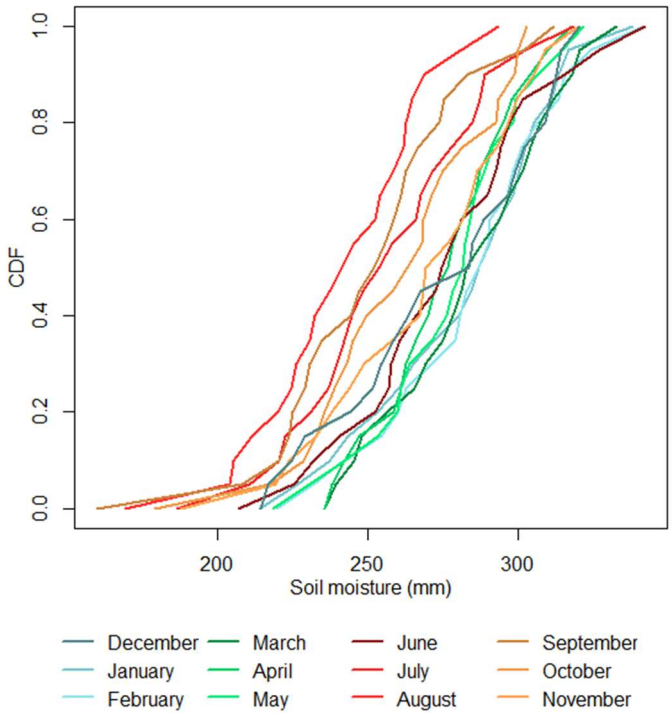


Figure 9: Monthly soil moisture CDFs at Austin, TX.

3.2.4.3 Texas Drought 2011 and California Drought 2015

Figure 10 shows two national soil moisture maps for September 2011 and May 2015. The map acknowledges the persistent dry condition in Texas during the 2011 drought, in which the vast majority of values were below the 0.05 percentile. During May 2015, Texas shows some recovery in soil moisture, even being wetter than usual but California and the west coast were experiencing low soil moisture values below the 0.05 percentile.

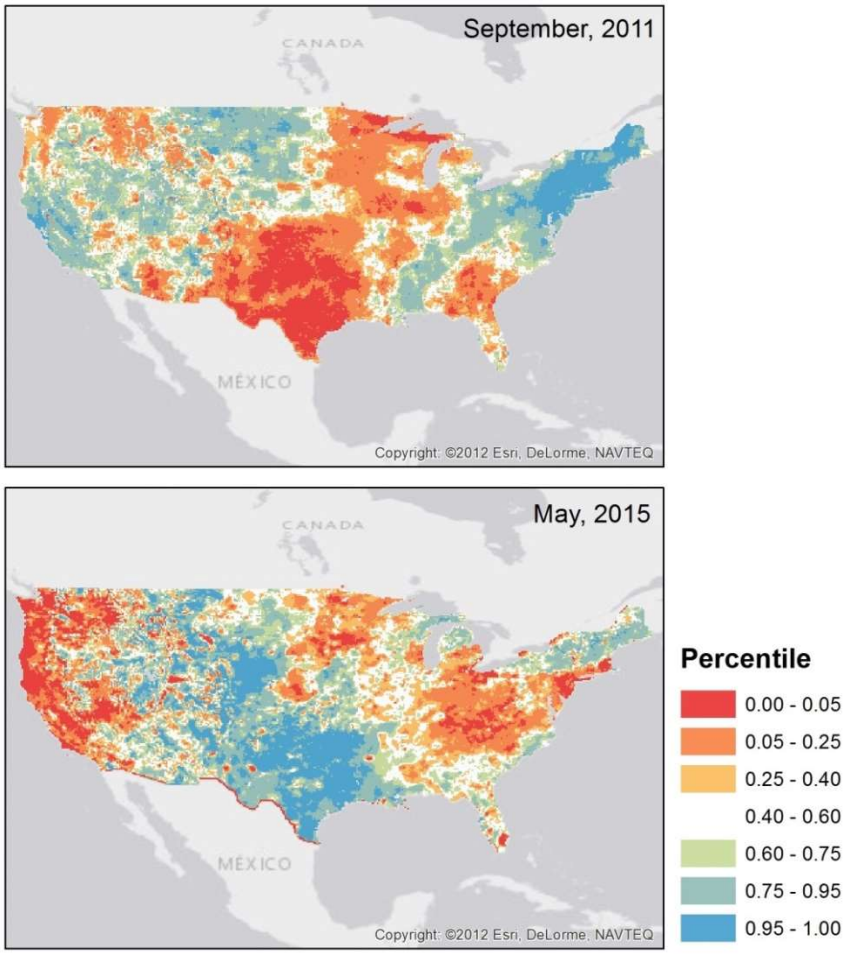


Figure 10: Percentile distribution of soil moisture in Texas. September, 2011.

Figure 11 is a map of the U.S. drought monitor (Miskus, NDMC, USDA, & NOAA, 2015) for September, 2011 and May, 2015. The areas with percentile values below 0.05 in Figure 10 are the areas marked as being subject of an exceptional drought in Figure 11. This suggests that the comparison of soil moisture values with the historical CDFs (in combination with other variables) might be useful for identification of areas experiencing drought conditions.

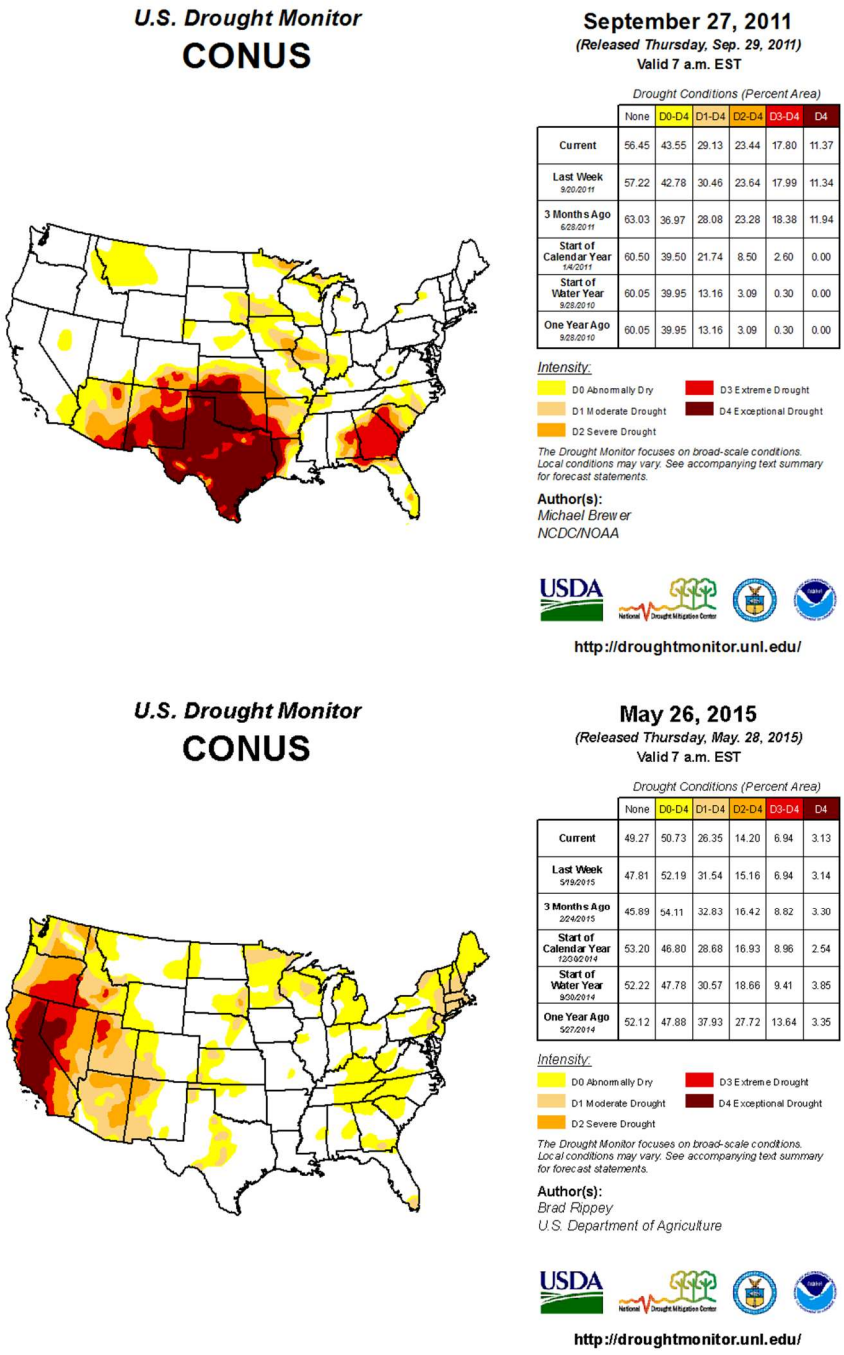


Figure 11: Classification of drought conditions from the U.S. drought monitor (Miskus et al., 2015) for September, 2011 and May 2015.

3.2.4.4 Halloween Flood, Onion Creek 2013

Figure 12 shows the soil moisture conditions at the Onion Creek watershed located south of Austin just before the flood on October, 31 2013. The left plot shows the soil moisture values in the previous 30 days before the storm, identifying a significant increase in the soil moisture due precedent storms. The plot on the right shows the CDF and the pre-storm soil moisture that corresponds to an 87 percentile. These wet conditions might have influenced a larger volume of runoff, for this case a more thorough analysis is needed.

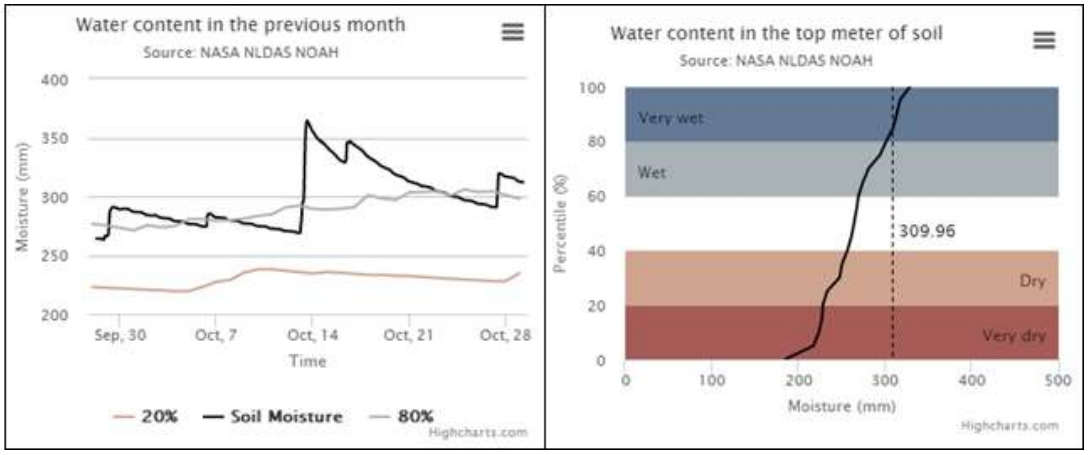


Figure 12: Pre-soil moisture conditions for the "Halloween storm" at Onion Creek on October 31, 2013. On the left, the variation of soil moisture, the 20 and 80 percentiles for the previous 30 days. On the right, the CDF distribution and the pre-storm soil moisture (dotted line).

Figure 13 shows the precipitation (top) and runoff (bottom) rates during the "Halloween Storm" at Onion Creek obtained from NLDAS. The plots show that around midnight on October 31, 2013 both variables peaked.

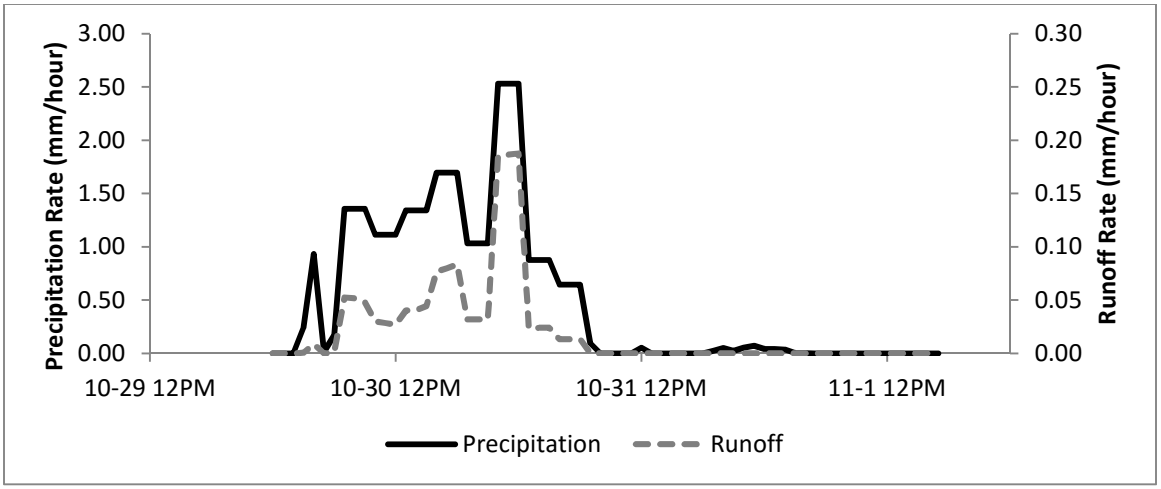


Figure 13: Precipitation and runoff rates (mm/hr) at Onion Creek during the "Halloween Flood" on October 31, 2013.

3.3 CDFS FITTING

The empirical CDFs were fitted by common probability distributions, using the calendar day or month statistics to estimate the parameters of the distributions. This process was repeated for all variables, all grid points in the United States, and all the calendar days and months. Therefore, distributions in which the parameters can be estimated efficiently were more desirable. Similarly to the empirical CDFs, the model distributions were fitted on a daily and a monthly basis. The common family distribution was the same for each variable but the parameters changed for each day (or month) and for each cell.

Table 3 shows the variables and the probability distribution models selected for the fits. The models for precipitation and runoff for the daily case, considered the two-step process in the empirical CDF calculation. The Bernoulli distribution models the probability of the event to occur (e.g. precipitation/no precipitation) and the Gamma distribution models the precipitation (or runoff) depth given that there was precipitation (or runoff).

Variable (X)	Time Interval	Model	Distribution	Y Distribution
Soil moisture	Daily/Monthly	$P(X)$	Normal	-
Temperature	Daily/Monthly	$P(X)$	Normal	-
Evapotranspiration	Daily/Monthly	$P(X)$	Gamma	-
Precipitation	Daily	$P(X\|Y)$	Gamma	Bernoulli
Precipitation	Monthly	$P(X)$	Gamma	-
Runoff	Daily	$P(X\|Y)$	Gamma	Bernoulli
Runoff	Monthly	$P(X)$	Gamma	-

Table 3: Variables and models of the probability distributions used to fit the empirical distributions.

In order to select the mentioned distributions, six common distributions were considered and tested at random points in Texas and for different calendar days: (1) Normal, (2) Gamma, (3) Pearson type III, (4) Weibull, (5) Pareto, and (6) Exponential. The Pearson type III distribution was discarded due its complicated automatic implementation and because it did not represent a higher improvement in the fit in comparison with the other distributions. Similarly, the Weibull and Pareto distributions were not selected because their parameters had to be estimated using a numeric solver which increased significantly the computing time (in some cases it did not find the parameters) and the fits performed inferior to the other distributions.

The gamma distribution was selected for the evapotranspiration, precipitation, and runoff variables due is great performance fitting the empirical CDFs, the direct calculation of its parameters, and to avoid negative values for the lower percentiles (a problem if using the normal distribution) which have no physical meaning. In addition,

the exponential distribution was considered a strong candidate for the precipitation and runoff variables but it consistently performed inferior to the gamma distribution and in some cases the fit did not passed the validation test (Section 3.4), the inferior performance might be explained due a poor fit for the lower percentiles which increased significantly the error. The soil moisture and temperature variables showed persistently a normal behavior. In part, the consistency of the results can be explained due the nature of the variables, hence a normal distribution was selected.

3.4 VALIDATION OF THE FIT

The model fits were validated using the Kolmogorov–Smirnov (KS) test. The KS was especially suitable for this application and has advantages from other hypothesis testing methods because (1) the test checks if the data (empirical CDF) has the same distribution as the tested theoretical distribution. (2) The test is applied on the empirical CDF and not only on two parameters (traditionally the mean and variance). (3) The test does not depend on a specific distribution and can be used for any given distribution. (4) The test it is based on the maximum distance between the theoretical and the empirical CDFs. The KS statistics captures properly if a theoretical CDF deviates greatly in any part from the empirical CDF. And (5) the test can be efficiently implemented to check all the fits without using large computational resources.

If $S_{x,d}\big|_{t_1}^{t_2}$ is the subset of the time series of the variable x in the calendar day or month d from the dates t_1 to t_2. And $O_{x,d}\big|_{t_1}^{t_2} = \{x: x(1), x(2), x(3), \dots, x(n)\}$ the ordered set of $S_{x,d}\big|_{t_1}^{t_2}$. The KS statistic is defined by Sager, 2010 (Equation 10) as the maximum difference between the theoretical CDF $\left(F_d(x)\right)$ and the empirical CDFs. In the two-sided hypothesis, the KS checks that the distribution is not greater (Equation 11) and not smaller (Equation 12) than the theoretical distribution. Thus, the null hypothesis in the

two-sided test is that the data behind the empirical CDF is from the theoretical distribution tested.

$$KS_n = \max(KS_n^+,\ KS_n^-) \quad (10)$$

$$KS_n^+ = \max_{1 \le i \le n} \left(\frac{i}{n} - F_d(x) \right) \quad (11)$$

$$KS_n^- = \max_{1 \le i \le n} \left(F_d(x) - \frac{i-1}{n} \right) \quad (12)$$

Where:

KS_n: Kolmogorov-Smirnov statistic for two-sided case

KS_n^+: Kolmogorov-Smirnov statistic for the upper-tail

KS_n^-: Kolmogorov-Smirnov statistic for the lower-tail

$F_d(x)$: Theoretical CDF for the variable x and the calendar day or month d.

i: index of the ith element in $O_{x,d}\big|_{t_1}^{t_2}$

n: number of elements in $O_{x,d}\big|_{t_1}^{t_2}$

The p-value depends on the distribution of the KS statistic and the number of elements n in the $S_{x,d}\big|_{t_1}^{t_2}$. Marsaglia et al., 2003 developed a robust method for calculating the p-value for $2 \le n < 16{,}000$ and 13 digit accuracy, which is suitable for the present research where $n = 35$. The code is available in the C language and it is also implemented in R by R Core Team, 2014. The use of this methodology for the p-value calculation does not represent a significant additional computational time.

A significance value of 0.05 was selected, in which the null hypothesis that the empirical CDF follows the proposed theoretical CDF is rejected for smaller p-values. A fit was considered good if the p-value is greater than 0.1; the fit was also considered acceptable if the p-value was between 0.05-0.1 and no other fit provided a greater value.

3.4.1 Results

3.4.1.1 Example: CDFs fits for Onion Creek

Figure 14 shows on the left the empirical CDFs and the theoretical CDFs, and on the right the quantile-quantile plots for the grid cell at the outlet of Onion Creek watershed (Latitude: 30.3125, Longitude: -97.6875) south of Austin for June 15. The calculation of this fits was automated using R, and can be replicated efficiently for each grid cell and each calendar day (or month) in parallel. The KS statistic and the p-value were also stored and reported.

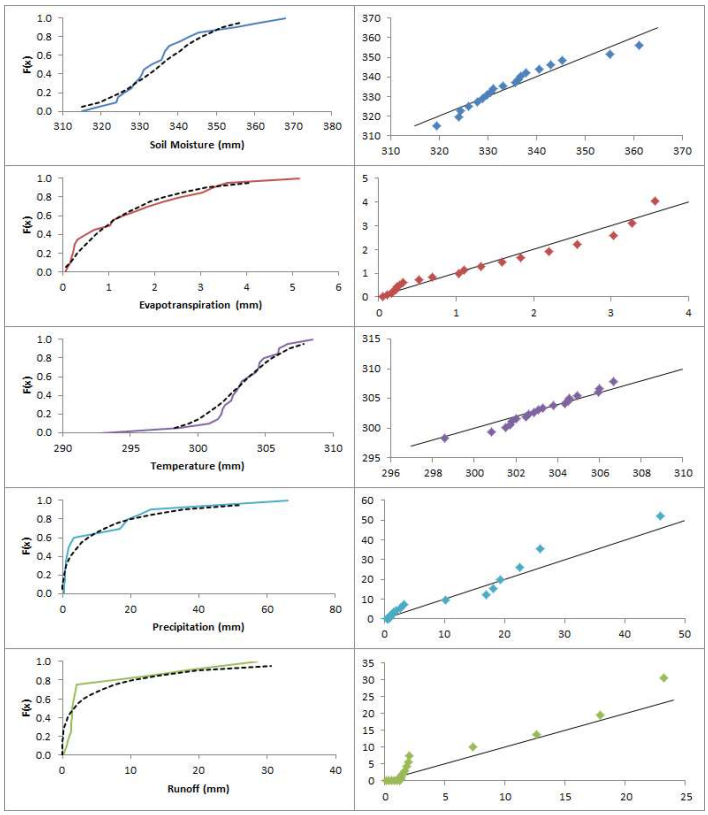


Figure 14: Comparision of theoretical versus empirical CDFs for the variables (top to bottom) soil moisture, evapotranspiration, temperature, precipitation, and runoff of the Onion Creek watershed on June 15. On the left, empirical (solid line) and theoretical (dotted line); on the right, quantile-quantile plots.

Table 4 shows the results of the fits in the previous figure (Outlet of the Onion Creek watershed, June 15). We can conclude that the fits pass the test with a significance level of 0.05 for the given p-values.

Variable	D statistic	P-value
Soil moisture	0.14	0.46
Evapotranspiration	0.18	0.21
Runoff	0.27	0.79
Precipitation	0.22	0.61
Temperature	0.15	0.36

Table 4: Kolmogorov-Smirnov results of the goodness of the fits for the outlet of the Onion Creek watershed on June 15.

3.4.1.2 Validation of the fits

The null hypothesis in the Kolmolgorov-Smirnov (KS) test states that the data behind the empirical CDF is from the theoretical (i.e. fitted) distribution tested. For a selected significance value of 0.05, smaller p-values indicate that the null hypothesis is false and the data does not come from the tested theoretical distribution. Large p-values show that there is no evidence that the data does not come from the theoretical distribution tested. Thus the results performed by the KS method are a weak form of validation if the sample size is small. For the current application and the 35 year period of data, the results from the KS test were sufficient enough to consider the fits validated.

Figure 15 is a summary of the validation of the daily fits per hydrologic region in the United States. A map and a list of the hydrologic regions are included in Appendix III. For all the hydrologic variables and regions the p-values were greater than 0.4 in at

least 75% of the fitted distributions (with the exception of the regions 16, 17, and 18 for evapotranspiration and some outliers in the runoff and temperature variables). At least 90% of the fitted distributions passed the test, and for the locations (in time and space) that did not passed the test, a distribution can be estimated from the adjacent fits.

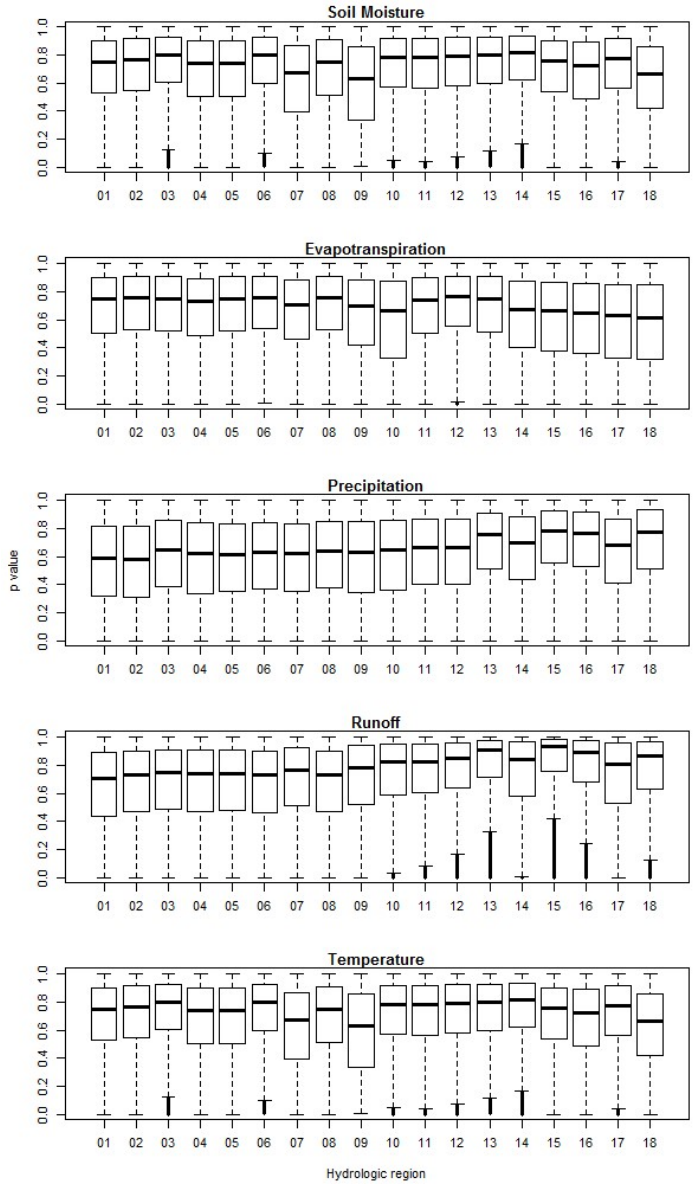


Figure 15: Results of the validation of the fits for each hydrologic variable and each hydrologic region in the United States.

Figure 16 is also a summary of the p-values of the daily fits but per month of the year. The plots also show that more than 75% of the tested fits passed the test with a p-value of 0.4 or greater, with a median value around 0.7.

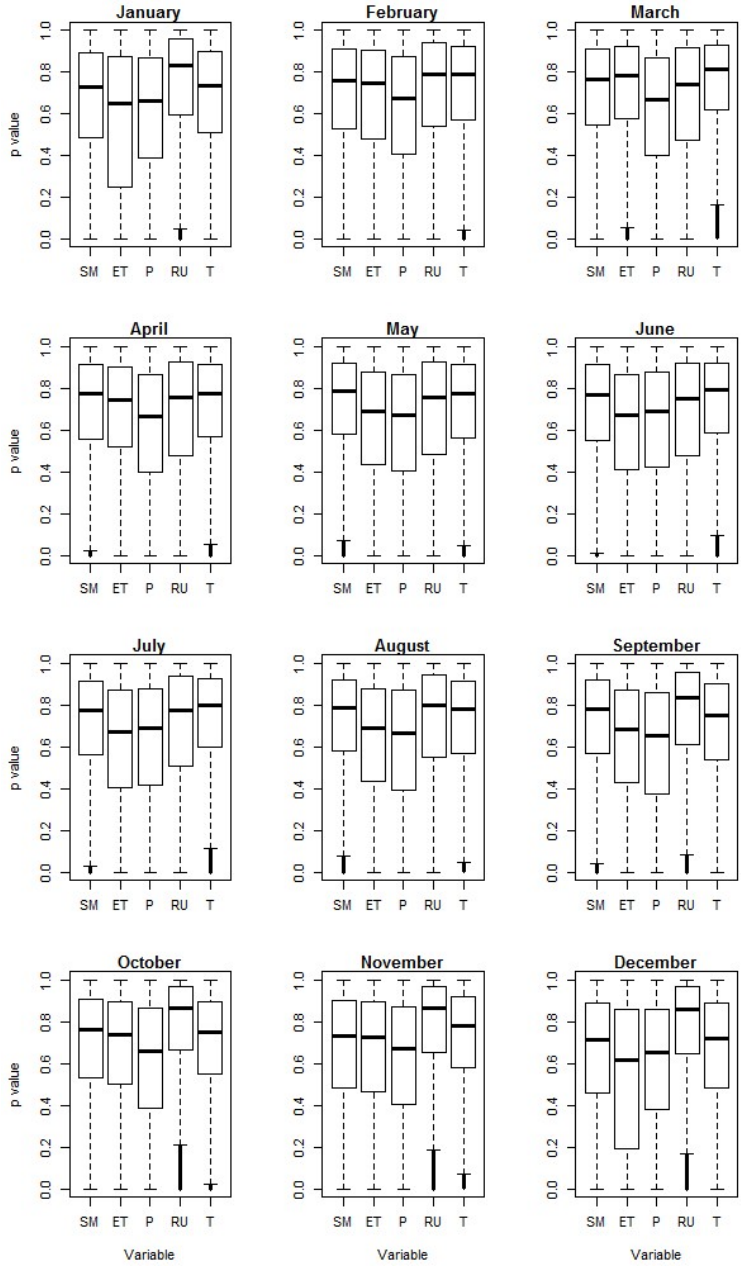


Figure 16: Results of the validation of the fits for each hydrologic variable and calendar month.

Figure 17 shows the fraction of the total fits tested per hydrologic variable for the range of p values in 0.1 increments. For all the variables the majority of the fits have a p value greater than 0.9. Only a small fraction of the fits did not passed the test (p value < 0.05) or barely passed it (0.05 < p value < 0.10). This fraction per hydrologic variable was: 0.03 for soil moisture, 0.06 for evapotranspiration, 0.06 for precipitation, .04 for runoff, and 0.01 for temperature. The small fraction of the modelled CDFs that did not passed the test can be derived from adjacent cells for statistical analysis without compromising the results in a similar procedure with zonal averages as described by Moody, King, Schaaf, & Platnick, 2008.

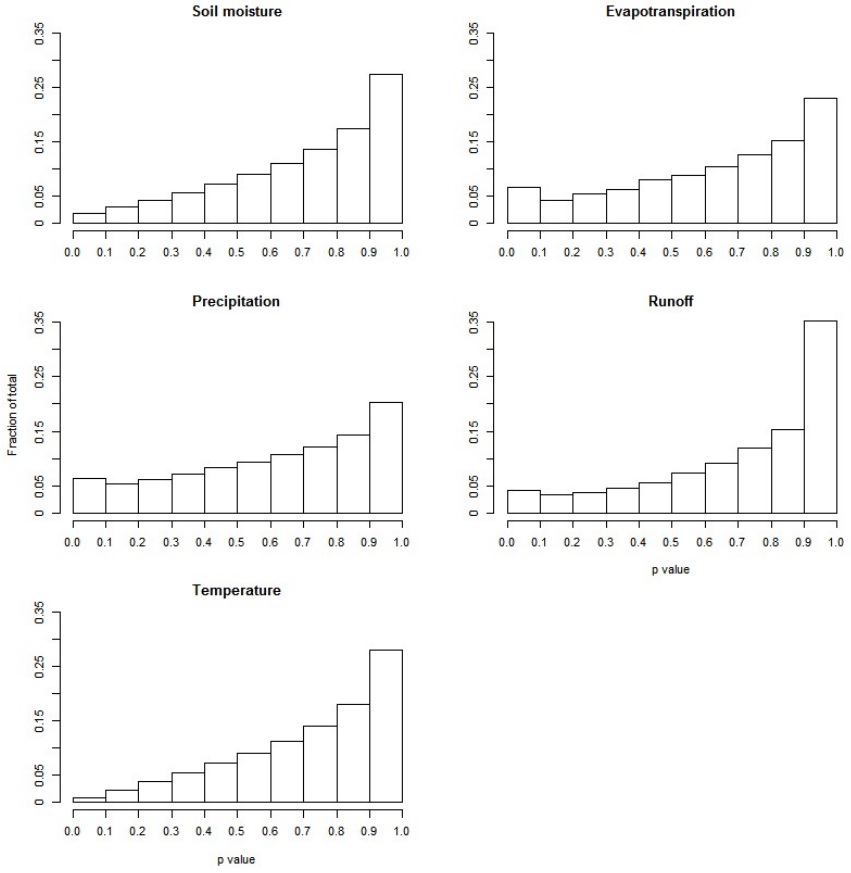


Figure 17: Fraction of the total fits that fall in the given p value range per hydrologic variable.

Figure 18 shows a density plot of the p values versus the mean parameter used in the fit. Darker blue cells represent areas where a combination of p value – mean is more common than in light blue areas.

The soil moisture and temperature plots show symmetric distributions of the p value in respect of the mean (ordinate axis), which confirms the selection of the normal distribution as appropriate. The distributions of the values are centered around 260mm and 12 °C respectively. The majority of the data pairs are located for large p values.

The plots are centered where most of de values occur (darker blue) although values outside the range showed in the abscissa axis. For the evapotranspiration, precipitation, and runoff plots the p value – mean pairs are centered close to a zero mean and decrease as the mean increases. The pair density is also higher for large p values as in the soil moisture and evapotranspiration variables.

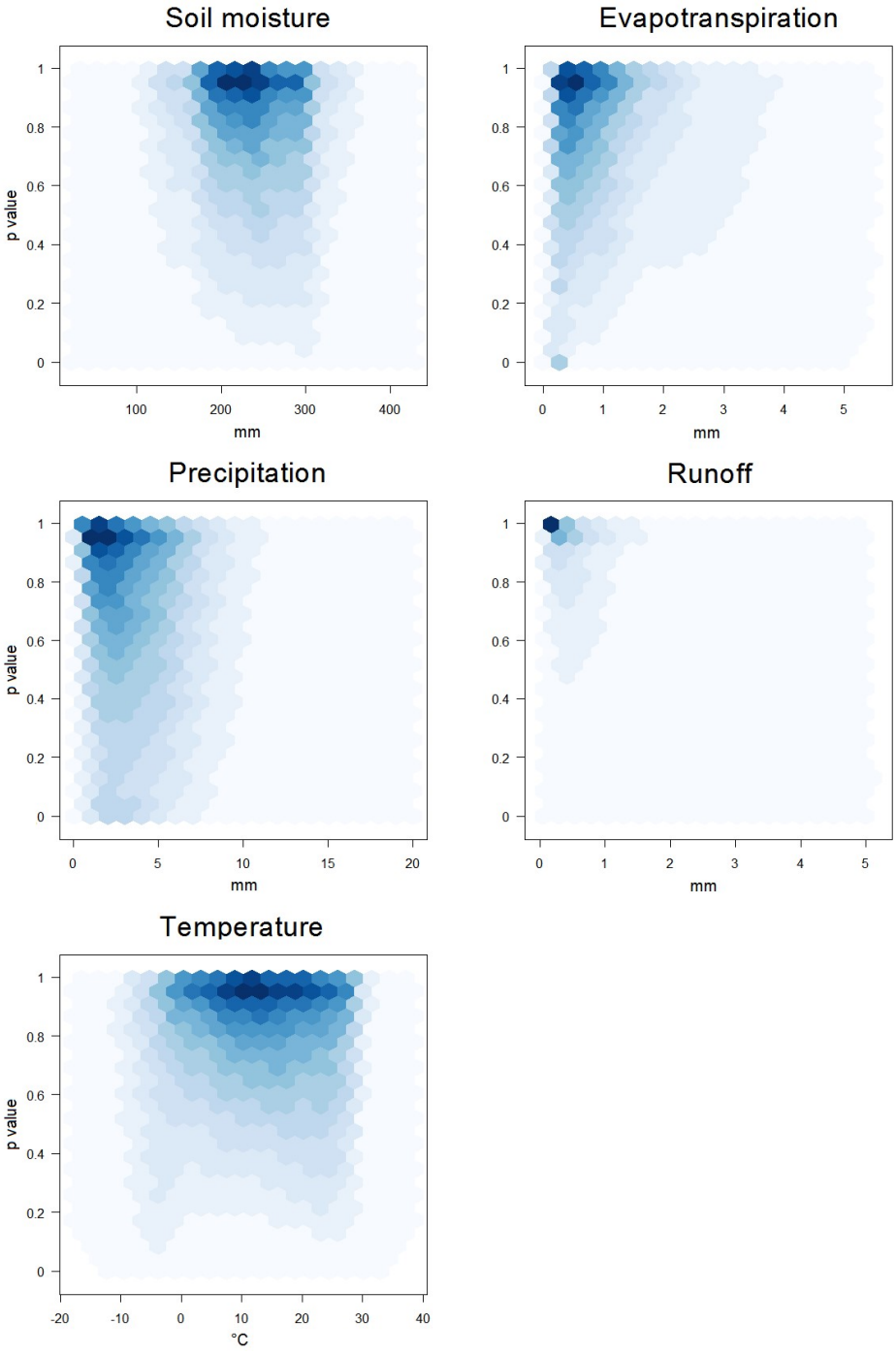


Figure 18: Density plot for the results of the fits for the combination of p values and mean. High density areas (darker blue) show more frequent values than low density areas (lighter blue).

It is noticed that there is a small bias in the evapotranspiration distributions in a small proportion of the fits (Figure 17and Figure 18) for the smaller evapotranspiration values. The fits perform constantly less satisfactory for the values close to zero. This is due NLDAS estimate negative evapotranspiration values and the gamma distribution enforces positive-only values. In general, the gamma distribution was the adequate for most range of values even though the fits did not perform as well in areas with low or negative values.

3.4.1.3 Modelled CDFs

The use of the fitted CDFs has advantages over the empirical CDFs:

- Data reduction: A pair of probability – value can be obtained for each hydrologic variable, for a given location and day or month of the year. The probability – value data is obtained using only the computed mean and standard deviation, suppressing the requirement of storing the complete hourly time series.
- Smoother functions: The fitted CDFs are smooth functions instead of jagged empirical calculations; they represent the historic trend without local fluctuations. Smoother and continuous CDFs functions are easier to use and implement in statistical modeling.
- Unbounded by historic minimum or maximum values: The fitted CDFs can associate a probability of occurrence to values that have not been present in the historic time series. Moreover, for the precipitation, runoff, and evapotranspiration variables, the gamma distributions impose a positive-only rule for the values. These two characteristics of the fitted

CDFs avoid unrealistic values and lead two more stable models when used in simulation or other applications.

- Representative of the historic conditions: The values are fitted using 35 years of data and validated the Kolmogorov-Smirnov test. Only a small fraction of the total number of fits did not pass the test (Figure 17) but an appropriated fit can be inferred by their close (in space and time) neighboring cells.

Figure 19 shows a comparison between the empirical and the fitted CDFs. The top plot shows the empirical CDFs (solid lines), the middle plot shows the fitted CDFs (dashed lines), and the bottom plot overlaps both CDFs. The difference between empirical and fitted CDFs is small, as is the case for large p-values (large majority of the cases). The fitted CDFs are smoother and are not influenced by local fluctuations.

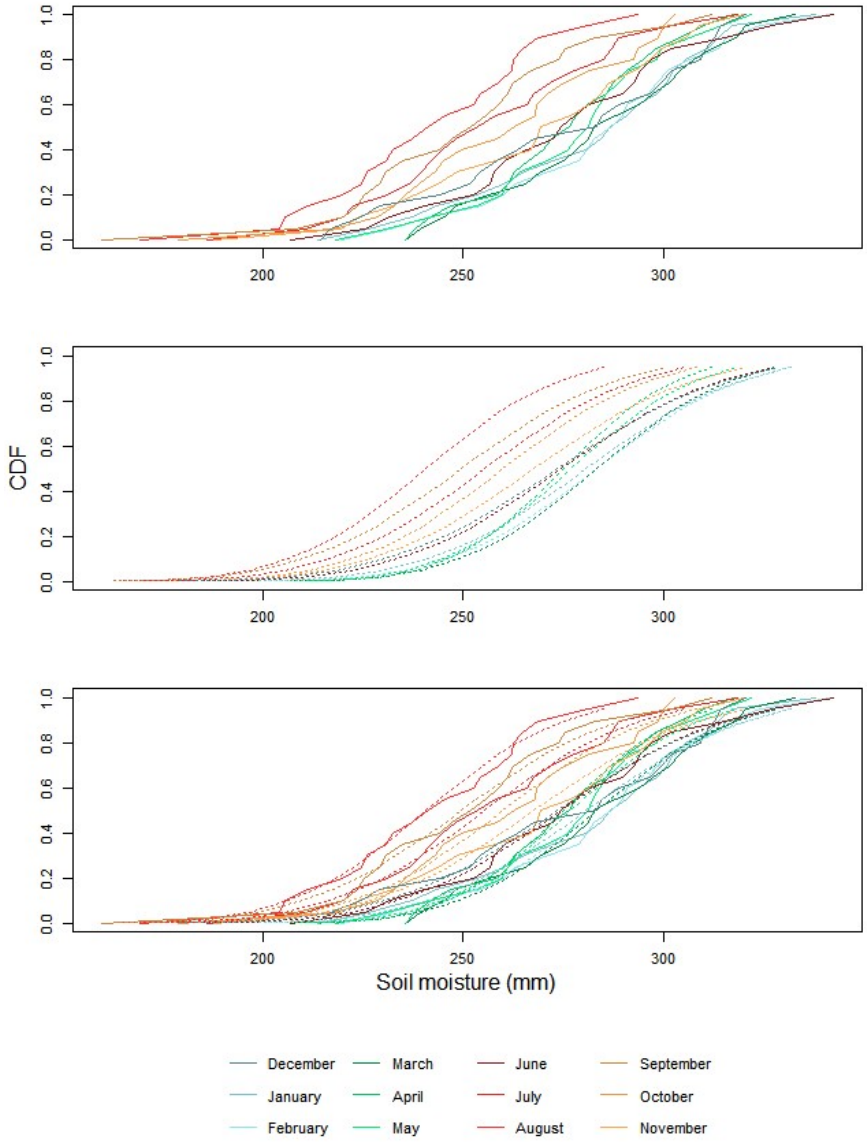


Figure 19: Monthly soil moisture CDFs at Austin, TX. The top plot shows the raw empirical CDFs (solid lines), the plot in the middle shows the fitted CDFs (dashed lines), and the bottom plot shows both overlapped. The difference between empirical and fitted CDFs is small and local fluctuations are smooth out.

3.5 IN-CLOUD STORAGE

The results of the analysis were uploaded to a cloud storage service (Microsoft Azure), which provides flexible approaches to access it through web services or web applications. The tables design was especially relevant to improve performance during the processes of uploading and retrieving data. Two tables were created per time interval (day or month) for a total of four tables. The tables (with minor modifications for the daily or monthly time interval) are:

- Results of the multidimensional statistical analysis
- Latest values in NLDAS

The first table stores (1) the grid cell code, (2) the statistics (mean and standard deviation), (3) the fitted models, and (4) the results of the fits (p-value). The second table is dynamic, and is updated every time new data is available. It stores (1) the latest results in the NLDAS model, (2) the percentile corresponding to this value, and (3) the anomaly of the values (defined as the number of standard deviations from the mean).

Figure 20 shows the general table design. The continental United States is divided in 88 partitions in order to improve performance during the process of querying and accessing data. The values can be retrieved given the geographic partition, grid cell code (quad), and (only for the results table) the day or month of the year.

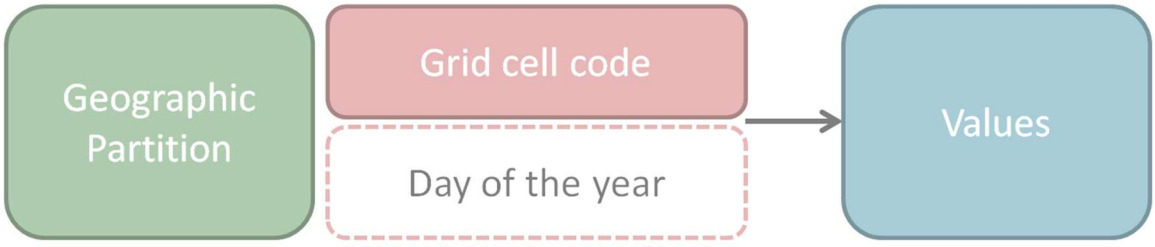


Figure 20: Tables design in cloud storage. The values are accessed by a geographic index, a grid cell code, and (optionally) the day or month of the year.

3.5.1 Statistical Analysis table storage

Table 5 shows examples of the values for the day interval in the cloud storage table. The United States is divided in 88 geographic partitions which are used as the *Partition Key* index. The *Row Key* column is a combination of the quad code representing the grid point (e.g. 30097-C6) joined by an underscore '_' to the four-digit calendar day MMDD (e.g. 0114 is January 14). For the month time interval, the grid point is joined to a two-digit calendar month MM (e.g. 04 is April). The table includes the mean, standard deviation, and the p-value results from the distribution fits for each hydrologic variable. The columns are named using the format: <*variable*>_<*parameter*> in which the variable options are: *SM*, *ET*, *P*, *RU*, and *T* for soil moisture, evapotranspiration, precipitation, and runoff respectively. The parameter options for all variables are: *mean*, s*tdev*, and *pvalue* for the computed mean, standard deviation, and p value respectively. An additional *prec* parameter is used for the precipitation and runoff variables in the daily case to retrieve the probability of a precipitation or runoff event.

Partition Key	Row Key	SM_mean	SM_stdev	SM_pvalue	…
81	30097-C6_0114	346.96	16.77	0.95	…
81	30097-C6_0115	347.27	16.33	0.94	…
81	30097-C6_0116	348.04	17.12	0.96	…
81	30097-C6_0316	345.67	15.42	0.51	…
…	…	…	…	…	…

Table 5: Example values of the soil moisture variable for the statistical analysis table.

The values stored in the azure cloud can be easily parsed using programing. A sample script of how to retrieve the information is shown in Figure 21. Three parameters

identify the cloud table: *account name*, *key*, *and table name*. Three parameters are used to get the statistical data: *PartitionKey* (obtained with latitude and longitud), *cell* (also obtained with latitude and longitude), and *mmdd* or *mm* which is the calendar day or month (time parameter).

```
>>> import azure.storage
>>>
>>> # Storage Parameters
>>> account = 'usnldas'
>>> key = '****************************************'
>>> table_name = 'DataValues'
>>>
>>> # Query parameters
>>> PartitionKey = '81'
>>> quad = '30097-C6'
>>> mmdd = '0114'
>>> table_service = azure.storage.TableService(account, key)
>>> entity = table_service.get_entity(table_name, PartitionKey, quad + '_' + mmdd)
>>> entity.SM_mean
346.96
>>> entity.SM_stdev
16.77
```

Figure 21: Sample python script to query and retrieve data from the statistical analysis table.

3.5.2 Latest Results in NLDAS table storage

Table 6 show examples of the soil moisture values in the Latest Results table. Similarly as in the Statistical Analysis table three parameters are used to query and retrieve the data: the *partition key* which are the geographic partitions used as index keys, the *row key* column that is the grid cell code (quad code), and the *Date* column that corresponds to the latest value on the model. The columns are named similarly using the format: <*variable*>_<*parameter*> in which the variable options are: *SM*, *ET*, *P*, *RU*, and *T* for soil moisture, evapotranspiration, precipitation, and runoff respectively. The

parameter options for all variables are: *value* and *anom*, for the latest value and the anomaly (i.e. number of standard deviations from the mean).

Partition Key	Row Key	Date	SM_value	SM_anom	…
87	25097-H3	10/24/2015	279.97	0.19	…
82	25097-H4	10/24/2015	162.12	-1.76	…
82	25097-H5	10/24/2015	71.92	-3.10	…
81	30097-C6	10/24/2015	378.85	1.52	…
…	…	…	…	…	…

Table 6: Example values of the latest result table in the cloud storage.

Similarly to the statistical Analysis table, the data Latest Results table storage can be easily queried and retrieved. Figure 22 shows a sample script that access latest results data for a given location.

```
>>> import azure.storage
>>>
>>> # Storage Parameters
>>> account = 'usnldas'
>>> key = '****************************************'
>>> table_name = 'LatestResults'
>>>
>>> # Query parameters
>>> PartitionKey = '81'
>>> quad = '30097-C6'
>>> table_service = azure.storage.TableService(account, key)
>>> entity = table_service.get_entity(table_name, PartitionKey, quad)
>>> entity.SM_value
378.85
>>> entity.SM_anom
1.52
```

Figure 22: Sample python script to query and retrieve data from the latest results table.

Chapter 4: Hydrologic Web Applications

Access to hydrologic data is critical for assessing extreme events assessment; it can help emergency responses and identify vulnerable areas. After hydrologic data is shared using the latest developments in information technologies, it can be used to expose and inform of current conditions or historic statistics through dynamic web applications. These dynamic web applications are an integration of data, maps, and web services founded in the use of standards.

Three web map applications were developed: (1) a web app showing the latest NLDAS soil moisture values in Texas and their comparison with the historic probability distributions, (2) a similar web app but for the continental United States and expanded for five hydrologic variables (i.e. soil moisture, evapotranspiration, precipitation, runoff, and temperature), and (3) a web application to explore, plot, and map hydrologic data called the Data Rods Explorer, this web app includes data from three datasets: LDAS (NLDAS and GLDAS), the Tropical Rainfall Measuring Missing (TRMM) (Simpson, Kummerow, Tao, & Adler, 1996), and the Gravity Recovery and Climate Experiment (GRACE) (Tapley, Bettadpur, Watkins, & Reigber, 2004). The first two web applications used the results of the statistical analysis in Chapter 3. The third web application is a wrapper, mapper, and plotter of the framework described in Chapter 5.

The goals of the web applications were: (1) being integrative platforms of web services including map and data services, (2) being successful examples on sharing the analysis of large hydrologic datasets though dynamic and reliable websites, (3) exposing the results in simple but robust web applications, (4) improve the meaning and facilitate the interpretation of hydrologic information, and (5) ease the data access and

visualization of hydrologic datasets. The goals were achieved through the incorporation of web services and technologies.

4.1 WEB APPLICATIONS ARCHITECTURE

The web applications architecture was structured with three (Figure 23) underlying components and implemented using two software alternatives: ArcGIS (Esri, 2015) and Tethys (Jones et al., 2014). ArcGIS is a complete and extensive software used in several geography fields: such as GIS, mapping, web GIS, and geoprocessing. Tethys is an innovative platform to facilitate the development and deployment of web applications in water resources. The Tethys platform includes ready-to-use mapping, data, and plotting templates that ease the process of displaying hydrologic data.

The architecture components for both platforms are:

1. Client-side
2. Server-side
3. Cloud

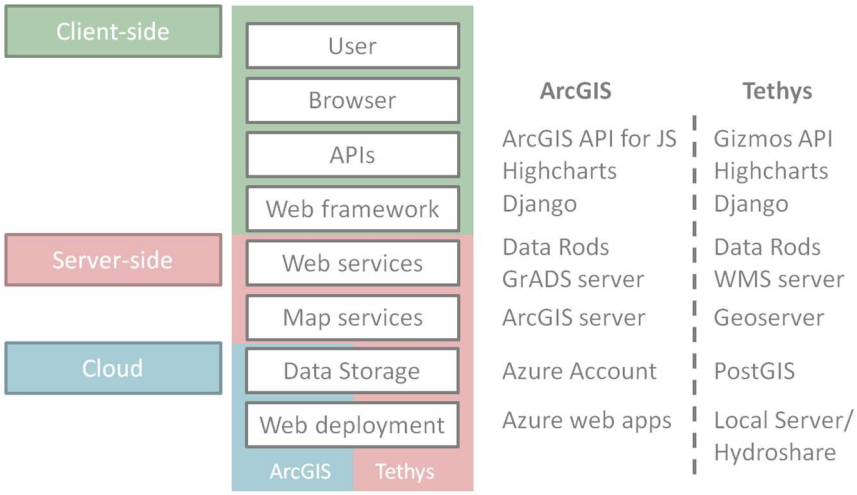


Figure 23: Underlying components of the web applications architecture: client-side, server-side, and cloud. The architecture was implemented using two software package alternatives: ArcGIS and Tethys.

The (1) client-side is the web application itself that displays the results of the statistical analysis or the NASA datasets through maps and charts. It is accessed by the users employing a browser, the only software requirement. The (2) server-side consists of two elements: map layers that contain the geographic information and the data coming from the GrADS, the WMS or the data rods web services. The (3) cloud component is where the application resides and where the temporal-spatial statistics from Chapter 3 are stored. The data stored in the cloud is shared as a web service and accessed using the Azure Software Development Kits (SDK).

In the ArcGIS system, information flows from one component to another using the Django python web app framework and an Application Program Interface (API) specifically developed for web mapping (ArcGIS API for JavaScript). Using the Tethys platform (Jones et al., 2014), the information is also connected using the Django web framework and the platform already includes a software suite and a API (Gizmos API) to facilitate the integration between elements.

4.1.1 Client-side

On the client-side is a web application accessed in a browser, in which the HTML is deployed by a Django website. Using the ArcGIS platform, the maps are loaded using the ArcGIS API for JavaScript. In contrast, the Tethys platform uses OpenLayers, which is an open-source mapping library (OpenLayers development team, 2015). For both platforms, the plots were created using the Highcharts JavaScript library (Highcharts developing team, 2015). The user interacts with the web application clicking on the map. The click starts an event that parses the location from the ArcGIS API for JavaScript or OpenLayers to Django. Django uses the location to get the LDAS grid code through a Python script. Additional Python scripts retrieve the data from the cloud storage or the

data rods web service. The data is returned to the HTML from Django to the ArcGIS API for JavaScript or OpenLayers and the Highcharts library for the pop-up and the plots.

Figure 24 shows how the data is displayed on the Texas soil moisture web application. The statistics for the calendar day, the latest value, the anomaly, and the corresponding percentile are presented on the pop-up. The CDF curve for the calendar day and the latest result (dotted line) are shown on the top-right chart. The bottom-right chart displays the previous thirty days from the data rods and the 20 and 80 percentiles.

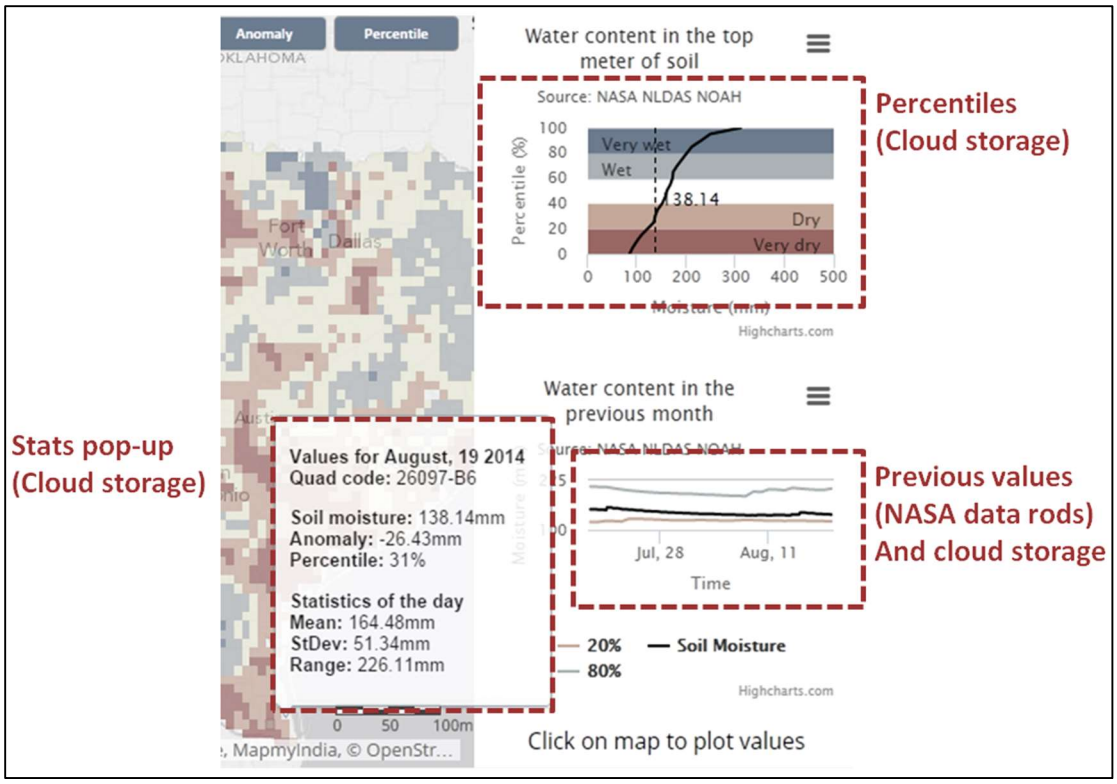


Figure 24: Data displayed in the Soil moisture web application. The data can be presented in different ways: pop-up, tables, and charts. And coming from different sources: cloud storage and the data rods server.

4.1.2 Server-side

Six different servers (Table 7) support the web applications: (1) the Data Rods server which is time-indexed and is used to retrieve hydrologic (i.e. LDAS, GRACE, or TRMM) data. The server is accessed by Django through python scripts and the data is parsed as a time series to the plots. (2) The WMS server that is used to retrieve the latest conditions to update the NLDAS statistical web app and to display the rasters in the Data Rods explorer web app. The server is also accessed by Django through python scripts and the geographic data parsed to a map using the ArcGIS API for JavaScript or OpenLayers. (3) The GrADS server was used similarly to the WMS server but in the Texas soil moisture web app. The WMS server proved to be more efficient given that the information is already georeferenced and in raster format instead of text files. (4) An ArcGIS server at the Center for Research in Water Resources (CRWR) that stores and publishes the layers with the latest results used in the NLDAS stats and the Texas Soil Moisture web apps. (5) A Geoserver included in the Tethys platform. It is used by the Data Rods explorer app to temporary store the rasters requested by the users. The server allows the maps to be displayed in the map using the OpenLayers library. And (6) Hydroshare, a server already setup to host Tethys apps. The Tethys platform requires to be deployed in a Linux server with a special configuration which is an additional effort to be implemented in the cloud.

Server	Type	Platform	Usage
Data Rods	Web service	ArcGIS/Tethys	Retrieve time series data given the location (time-indexed).
WMS	Web service (map)	Tethys	Retrieve raster files for a given time (space-indexed).
GrADS	Web service	ArcGIS	Retrieve text files with data for a given time (space-indexed). Used in the Texas soil moisture web app and replaced by the WMS server in the NLDAS stats web app.
ArcGIS (CRWR)	Map service	ArcGIS	Contains the layers with the latest results, the computed anomalies, and percentiles.
Geoserver	Map service	Tethys	Stores the rasters being displayed in the Data Rods explorer web app.
Hydroshare[3]	Web deployment	Tethys	A place to deploy the Data Rods Explorer web app, alternative to the Azure websites in the cloud.
PostGIS[4]	Storage	Tethys	A place to store data (e.g. statistics) alternative to the cloud.

Table 7: List of servers used by the web applications

[3] At the moment of writing the Data Rods Explorer web app is not published on the internet. It will be available as part of the Tethys apps showcase (http://demo.tethysplatform.org/).

[4] The web applications do not use PostGIS but it is included in the Tethys platform and is an alternative place to store data instead of the Azure Storage in the cloud.

4.1.3 Cloud

The cloud is used in the Texas soil moisture and the NLDAS statistical web apps. It is implemented using the Azure cloud from Microsoft. It includes two parts: web deployment and data storage. The web deployment component is the location of (1) the website itself (i.e. html files and images) deployed using the Django framework, (2) the python libraries, functions, and scripts, and (3) JavaScript code for mapping and plotting. The web deployment on the cloud makes available the two web apps in the internet to the public, through the urls: http://texassoilmoisture.azurewebsites.net and http://statsnldas.azurewebsites.net respectively. The data cloud storage contains the tables described in Section 3.5. The data is accessible through the Django application itself using the Python Azure SDK as shown in Figure 21 and Figure 22.

Table 8 shows the pros and cons of storing data or deploying the websites in the cloud or on a local server. Both technologies are good for storing information online, its selection in a specific project will depend on the specific characteristics of it.

Technology	Cloud		Server	
	Component	Pros and Cons	Component	Pros and Cons
Storage	Azure Storage Account	Easy accessible through the Python SDK and scalable. The cost increases with the number of users and the total storage.	PostGIS	Integrated with the Tethys platform and uses the PostgreSQL relational database. Additional effort is required for setting up the system and maintenance.
Web deployment	Azure web apps	Ready to deploy, Automatic assignation of a domain name, scalable. Higher cost	Local server	Upgrading the system requires physical modifications, maintenance require expertise. Lower cost

Table 8: Comparison of cloud and server technologies for data storage and web deployment.

4.2 SOFTWARE PACKAGES ALTERNATIVES

The ArcGIS and the Tethys platform were proven to be adequate solutions for the development, implementation, and deployment of hydrologic web applications. For the present research, both solutions were used. The selection of the most efficient solution depends on the volume of users, the local resources, and a cost analysis. It is worth

mentioning that the Tethys platform can be deployed on a virtual machine in the cloud, although the size of the machine can increase the total cost and therefore reduce the cost-benefit relationship of the solution.

4.2.1 ArcGIS platform

The ArcGIS platform has the advantages of: (1) having a large development team behind the software, (2) it is highly maintained and upgraded, (3) it is the leading software in GIS in the industry, (4) it is extensible well documented, and (5) it provides technical support. The disadvantages are: (1) it is commercial software, (2) the APIs for web mapping are rapidly evolving, and updating these applications can be demanding.

4.2.2 Tethys platform

The Tethys platform has the advantages that: (1) it is based on open-source solutions (i.e. Geoserver, PostGIS, and OpenLayers), (2) it is a wrapper for web development, common solutions (e.g. mapping and plotting) are already implemented in the Gizmos API and are easy and quick to use, (3) the documentation is satisfactory, (4) it is evolving and it is increasing in functionality. The disadvantages are that: (1) there are not enough users outside the research team at BYU or academia, and (2) maintenance and the future of the project depends on external funding.

4.3 RESULTS

4.3.1 Soil moisture map for Texas

The Texas soil moisture web app (http://texassoilmoisture.azurewebsites.net) displays the latest soil moisture values of the NLDAS-Noah model and their comparison with historic values. It includes three layers: (1) the volume of water present as soil per unit area (i.e. millimeters or equivalent water depth), that is shown in absolute terms

(Figure 25). (2) The anomaly of the soil moisture values defined as the difference from the long term mean for the displayed day (Figure 26). And (3) the corresponding percentile of the soil moisture value obtained from the empirical CDF for that particular day of the year (Figure 27).

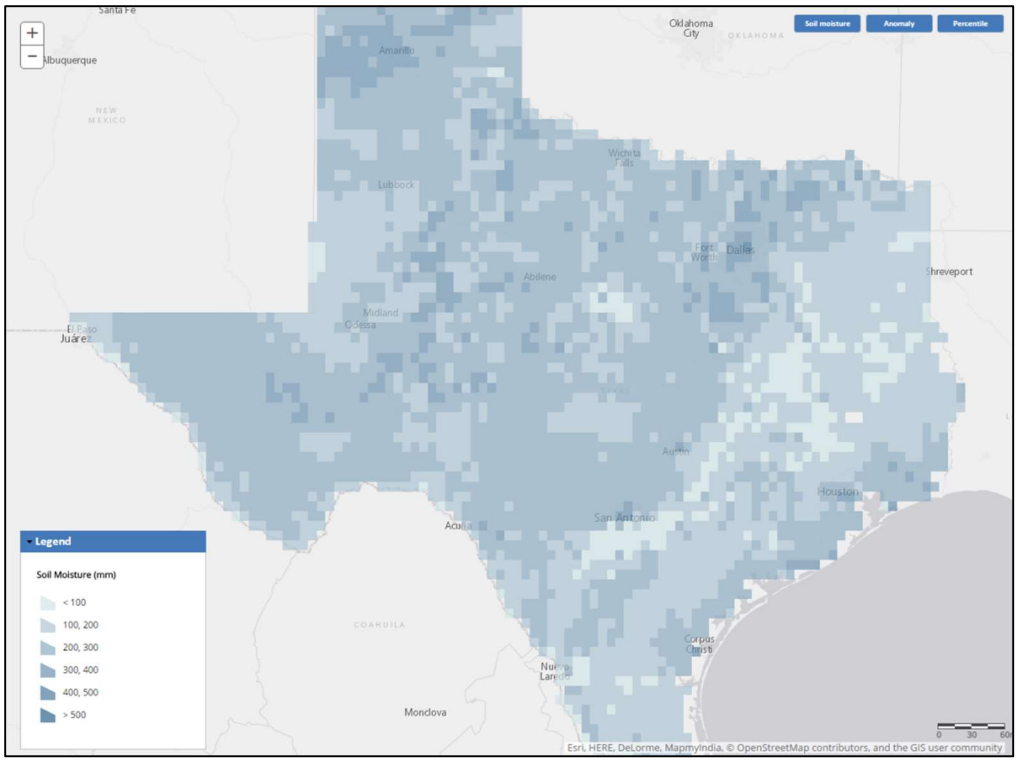


Figure 25: Layer of the soil moisture values in Texas (millimeters) on October 23, 2015. The areas with greater water equivalent depth (dark blue) are distinguished from the areas were its lower (light blue).

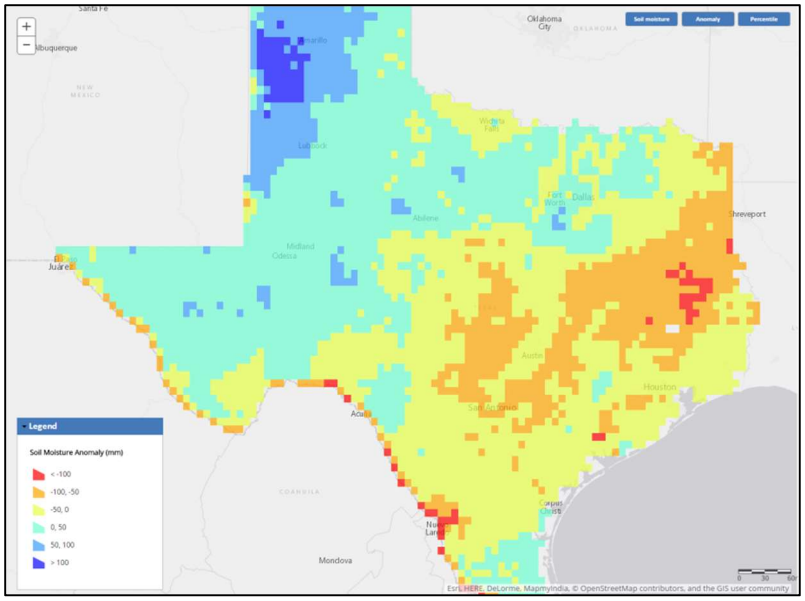


Figure 26: Layer of the soil moisture anomaly in Texas (millimeters) on October 23, 2015. The values are the difference from the mean. Negative values (warm colors) indicate that the current soil moisture is below the long-term mean, positive values (cold colors) indicate that they are above the mean.

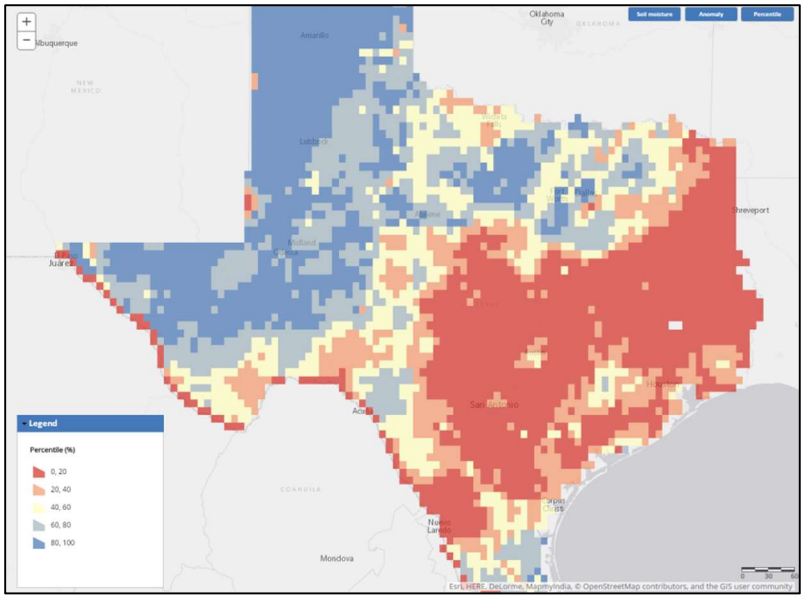


Figure 27: Layer of the soil moisture percentiles in Texas on October 23, 2015. The percentiles are obtained from the empirical CDFs of the daily time series (1979-2013). Dry areas (red) have a value under the 20 percentile. Wet areas (blue) have a value above the 80 percentile.

The ArcGIS server provides the map layers. The map layers are used to identify the geographic locations to a grid code in the statistical analysis and include the latest results for each cell. The maps also serve as an intermediate platform between the user and the cloud storage. The NASA data rods server provides the latest information and the time series data. It is accessed every time a user clicks on the map. The real-time response of the web app allows the charts to be drawn instantaneously as the user click on the map. The interaction between the geographic location clicked and the data retrieval from the statistical table was improved in the NLDAS statistical web app (Section 4.3.2).

Figure 28 shows the web app layout, the (1) top ribbon contains the name of the app, the date displayed (latest update), and the contact information. The (2) map shows one of the three layers which can be switched with the (3) controllers on the top-right corner of the map. After the map is clicked, a (4) pop-up with the statistics is displayed. The (5) top-right chart compares the latest value of the point clicked to the historic CDF and the (6) bottom-right chart show the variation in soil moisture from the previous 30 days and the correspondent 20 and 80 percentiles.

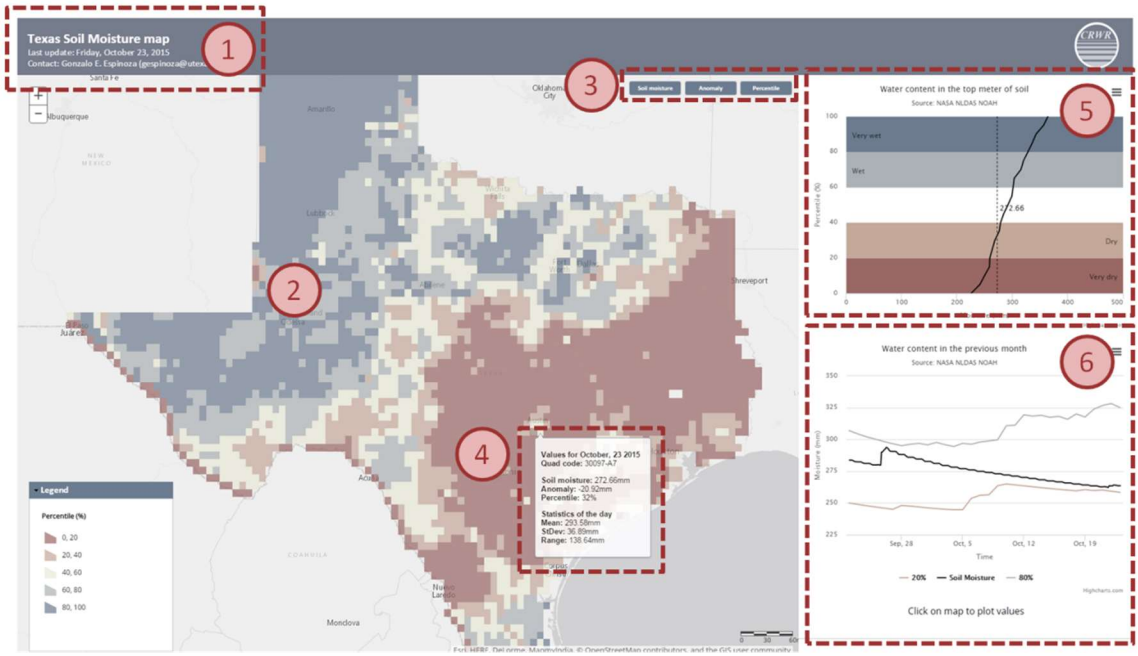


Figure 28: Layout and components of the soil moisture we app. (1) top ribbon, (2) map, (3) layer controllers, (4) pop-up with statistics, (5) plot of the CDF and current value of the day, and (6) previous values and their percentiles.

The Texas soil moisture web app was the first hydrological web application developed. The scope of the app was expanded to include more variables and all the grid cells in the continental United States, resulting in the NLDAS statistical web app. The two main changes were the use of the WMS server instead of the GrADS server and the use of image services (i.e. rasters, ArcGIS Resources, 2015) instead of using feature classes (polygons).

The use of the WMS facilitates the process of updating the layers with the latest information. The use of image services improves performance significantly, avoiding the need of drawing each grid cell (polygon) independently. The web app can handle more layers at the time (ten instead of three) and cover a greater area in a more efficient way.

4.3.2 NLDAS statistical map for the continental United States

The NLDAS statistical web app (http://statsnldas.azurewebsites.net/) displays the latest results of the NLDAS-Noah model for five variables: soil moisture, evapotranspiration, precipitation, runoff, and temperature. The web app includes ten layers, two per variable. One of the layers is for the absolute value of the latest results and a second layer is for the variable anomaly. Different from the Texas soil moisture web app, in the NLDAS statistical web app, the anomaly is normalized by the daily mean. In this case, the anomaly is defined as the number of standard deviations from the mean. A negative anomaly means that the current values are lower than the average daily conditions and positive anomaly values mean that they are above the average conditions.

The NLDAS statistical web app is an upgrade from the Texas soil moisture web app although the layout (Figure 29) remains the same (Figure 28). The components that were upgraded are: (1) the use of the WMS server instead of the GrADS server, which includes geographic references and allows a smoother updates of the latest conditions of the layers at the CRWR ArcGIS server. (2) The use of images services (i.e. rasters) instead of feature classes (i.e. polygons) which was a major improvement on user performance on the client-side of the web app, discarding the need of drawing each polygon independently. (3) The percentiles for extreme values are represented in narrower zones, improving the graphic communication of the meaning of extreme events. And (4) the cartography for in the anomaly layers is greatly enhanced due normalization. The layers easily exhibit the areas were extreme hydrologic conditions are present.

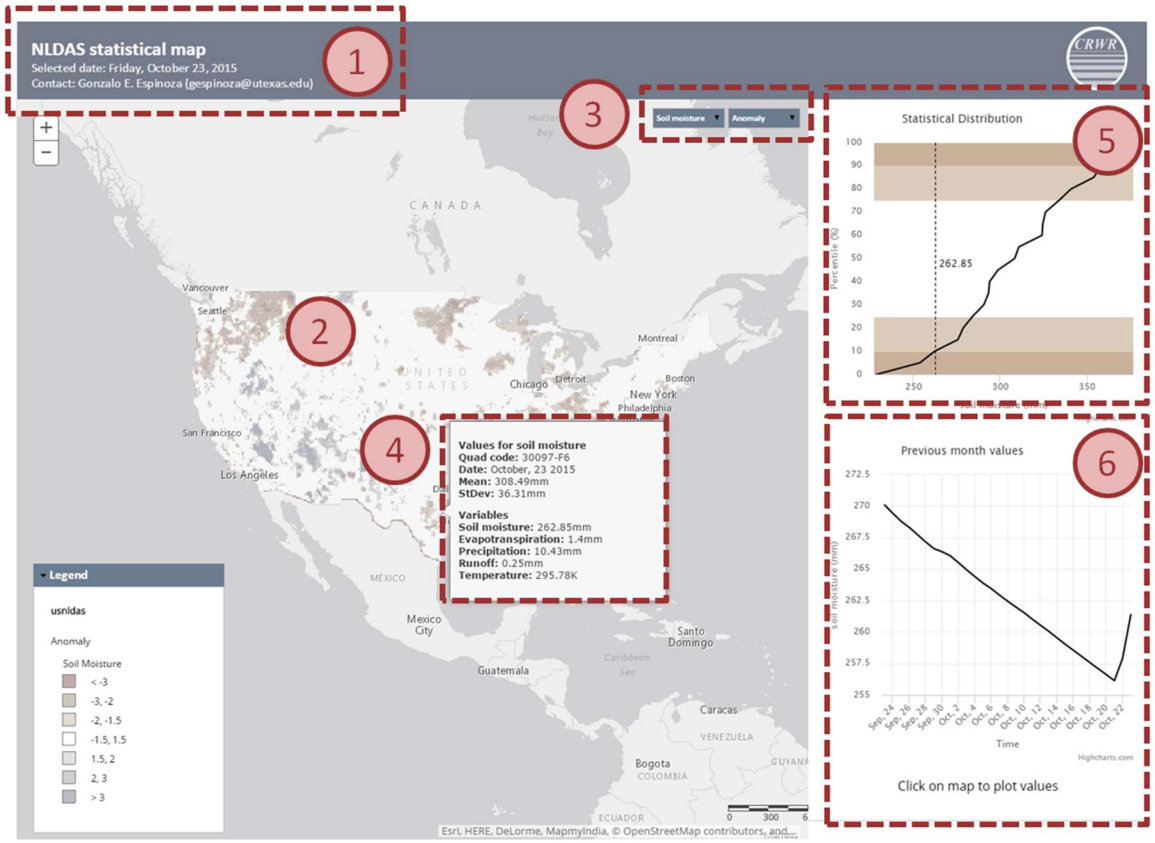


Figure 29: Layout and components of the NLDAS statistical we app. (1) top ribbon, (2) map, (3) layer controllers, (4) pop-up with statistics, (5) plot of the CDF and current value of the day, and (6) previous values.

4.3.2.1 Example: Storms on October 23, 2015

The NLDAS statistical web app captured the storms, the values, and their anomalies for the five variables across the country that occurred on October 23, 2015. The soil moisture layers (Figure 30) show the dryer regions (orange) in the Southeast and the Northwest. Also, regions were wetter (blue) soil moisture values were present.

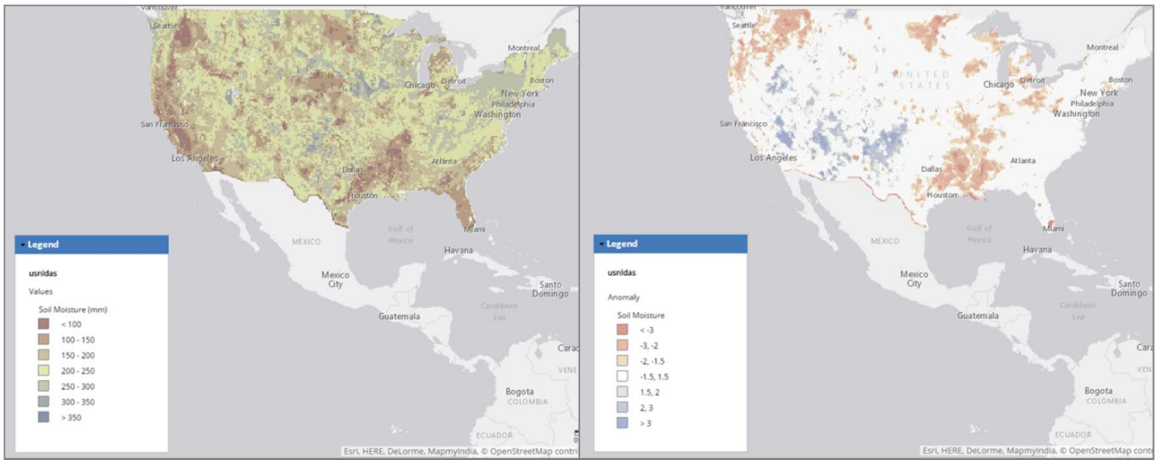


Figure 30: Soil moisture (left) and soil moisture anomaly (right) on October 23, 2015. Areas that were dryer (orange) or wetter (blue) than usual are identified by an anomaly value around three standard deviations.

Figure 31 displays the CDF (left) and previous values (right) plots on the Statistical web app for Austin, TX. The model predicted 330mm of soil moisture in the top meter. This value is around the 20th percentile of the daily CDF distribution. The bottom plot shows how soil moisture increases rapidly due rain events (on September, 26) and how it slowly decreases with time (from September, 26 to October, 20).

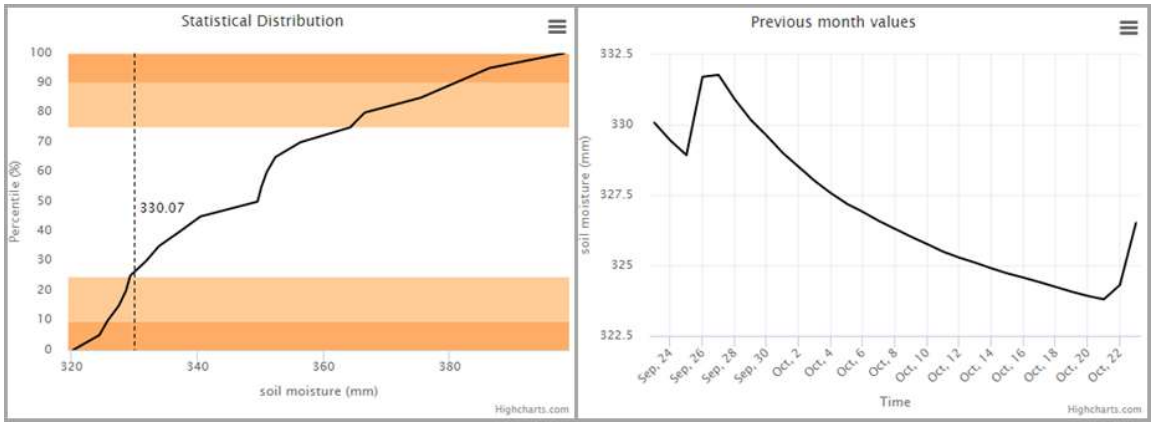


Figure 31: Plots of the daily CDF (left) and previous values (right) of soil moisture for Austin, TX on October 23, 2015.

Figure 32 shows the evapotranspiration (left) and its anomaly (right). The anomaly figure shows where the daily evapotranspiration values were close to the average conditions (white) or where they were above (blue) or below (orange) them. The spatial distribution is clearly identified, with larger anomaly in the central part of the country and smaller anomaly in the southeast.

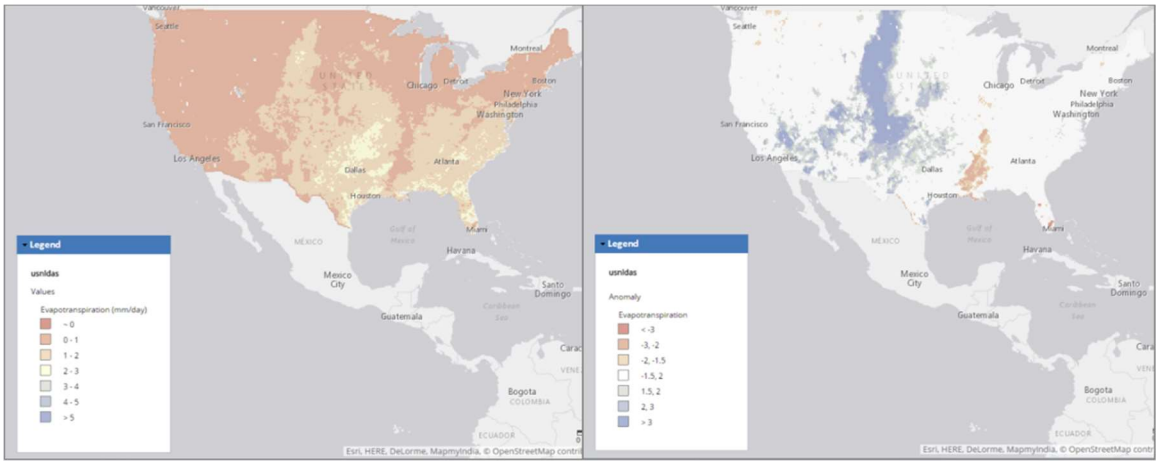


Figure 32: Evapotranspiration (left) and evapotranspiration anomaly (right) on October 23, 2015. The middle part of the country registered larger (blue) evapotranspiration values than the historic ones and the region close to the lower Mississippi river had lower (orange) evapotranspiration than the historic values.

Figure 33 and Figure 34 show respectively the spatial distribution of the precipitation and runoff depths (left) and their anomaly (right). The anomaly plots highlight the areas where the depths exceeded the historical conditions (blue) that in some cases was greater than three times the daily standard deviation. Further flooding studies can be focused on these regions.

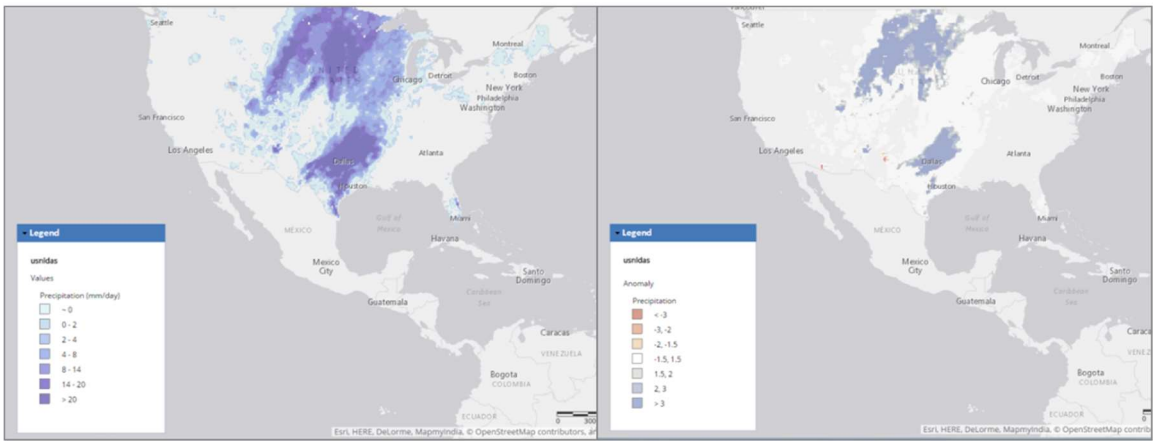


Figure 33: Precipitation depth (left) and precipitation anomaly (right) on October 23, 2015. The areas were the precipitation depth were statistically larger than the expected values (blue areas on the anomaly figure) are identified from areas were the precipitation depth is around the expected value (white areas on the anomaly plot).

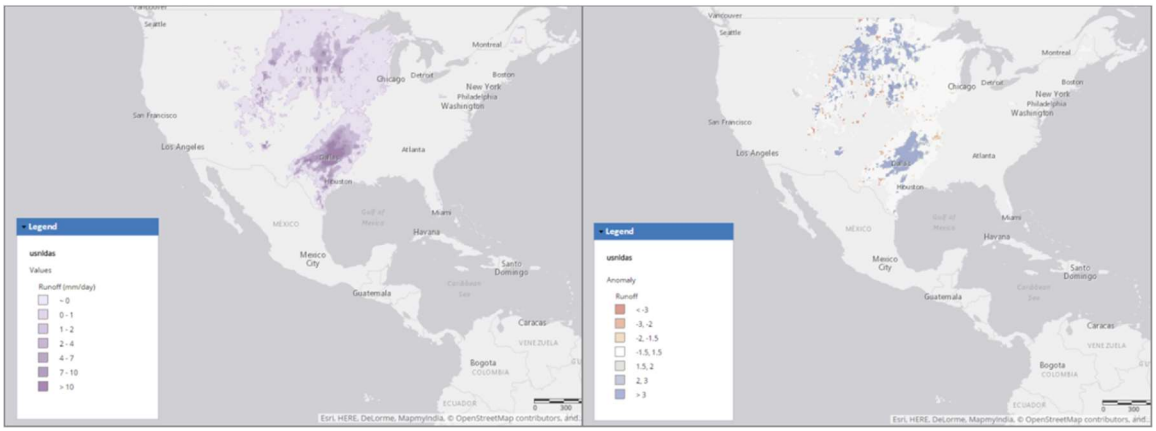


Figure 34: Runoff depth (left) and runoff anomaly (right) on October 23, 2015. The anomaly plot shows the areas subject of flooding (blue) were the runoff depth is statistically larger than historic conditions.

Figure 35 displays the daily temperature (left) and its anomaly (right). The northwest region was experiencing colder (blue) temperatures than the daily average. In contrast, the southeast was having warmer (orange) conditions.

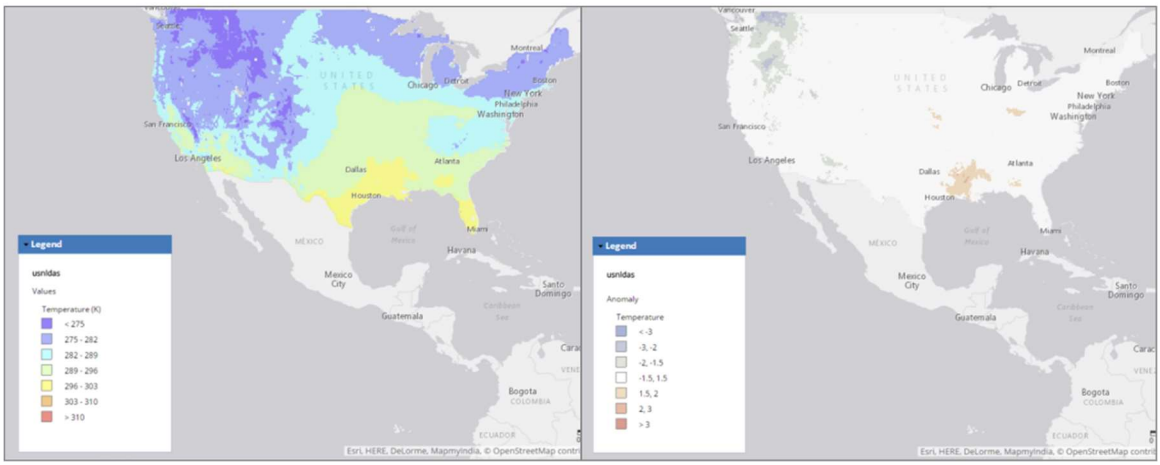


Figure 35: Temperature (left) and temperature anomaly on October 23, 2015. The anomaly plot shows that the northwest part of the country experienced colder (blue) conditions than the historic ones and that the region close to the lower Mississippi River (orange) was statistically warmer than usual.

4.3.3 NASA Data Rods Explorer

The Data Rods Explorer is a web app that allows the users to quickly obtain, plot, and map hydrologic data. The datasets available are: (1) LDAS for the North-American (NLDAS) and the global (GLDAS) Noah models. (2) The Tropical Rainfall Measuring Missing (TRMM) which provides global three-hourly precipitation data, and (3) the Gravity Recovery and Climate Experiment (GRACE) for three soil moisture percentiles data products: surface, root zone, and ground water.

The web application is different from the previous two, due the use of the Tethys platform instead of the ArcGIS platform. The design of the Data Rods explorer was focused on the user experience and to be interactive. The layout consists of six components (Figure 36):

- The GET parameters
- The top ribbon
- The main parameters selection

- The time series selection
- The map and plot container
- The bottom ribbon.

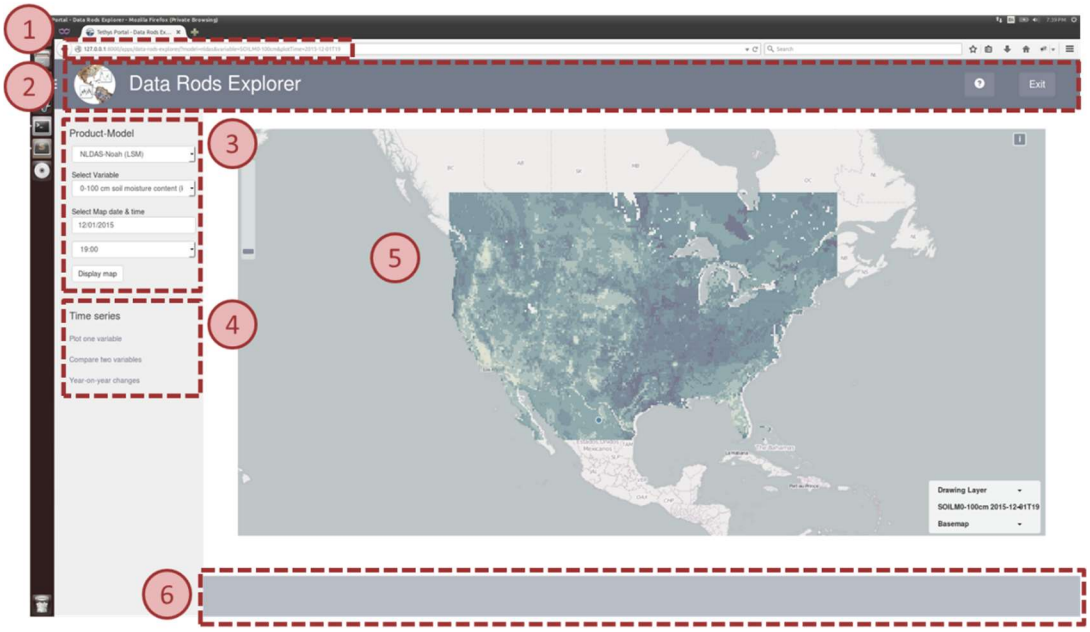


Figure 36: Data Rods Explorer web app layout: (1) GET parameters, (2) top ribbon, (3) main parameters selection, (4) time series selection, (6) map and plot container, and (7) bottom ribbon.

The GET parameters in the URL specify the options for the map and the plots. The top ribbon consists in a button to hide the left panel, the name of the application, and an exit button. The main parameters consist in three selection elements: the model, the variable, and the data and time; and one 'display map' button. After the button is clicked, a raster for the three selected parameters is loaded into de map. The raster comes from the WMS server and is stored temporarily (during the user session) as a layer in Geoserver. The time series component is for plotting the data using the Data Rods server. The three options are: (1) plot one variable, (2) compare two variables, and (3) year-on-year changes.

The *plot one variable* option plots the time series for the variable selected given the time interval (i.e. initial and final times). The *compare two variables* option allows the user to select a second model and variable and plot the two variables selected for a given time interval. The *year-on-year changes* option plots complete years of data for the variable selected which allow comparison between wet and dry years and the magnitude of the difference. The map and plot component is a container for the OpenLayers map and the HighCharts plots. Lastly, the botton ribbon displays messages that inform the user about the process run.

All the selection options (model, variable, and date and time) plus the time series parameters (e.g. start time, second variable, etc.) change dynamically the GET parameters in the URL as the options change. This allows the users to programmatically change the parameters and save the URL of their parameters of interest. Figure 37 has a sample URL for accessing the Data Rods Explorer web app proving the parameters for the home page. If the parameters are missing, default values are set.

```
http://127.0.0.1:8000/apps/data-rods-explorer/?          -Base URL
model=nldas                                              -Model
&variable=SOILM0-100cm                                   -Variable
&plotTime=2015-12-01T19                                  -Map date and time
```

Figure 37: Data Rods Explorer URL and main parameters: model, variable, and map date and time.

Figure 38 shows the changes in the base URL and the additional GET parameters required for the three time series options. The additional parameters are populated automatically when they are chosen or with default values if they are missing or erroneous. Appendix IV lists the models, their variables, and their respective codes used as GET parameters available in the Data Rods Explorer.

Plot One Variable

```
http://127.0.0.1:8000/apps/data-rods-explorer/plot?          -Base URL
&startDate=2015-01-01T00&startDate=2015-12-01T23             -Time interval
```

Compare Two Variables

```
http://127.0.0.1:8000/apps/data-rods-explorer/plot2?         -Base URL
model2=trmm                                                  -2nd Model
&variable2=precip                                            -2nd variable
&startDate=2015-01-01T00&startDate=2015-12-01T23             -Time interval
```

Year-on-Year Changes

```
http://127.0.0.1:8000/apps/data-rods-explorer/years?         -Base URL
&years=1990-1995,2012,2015                                   -Selected years
```

Figure 38: Time series options, changes in the base URL and additional GET parameters required: base url (blue), time interval (green), secondary model (red), and secondary variable (purple)

Figure 39 shows an example of the Data Rods Explorer interface after a raster is loaded. The map shows the NLDAS-Noah output of the top meter soil moisture for Texas on December 3, 2015 16:00 UTC. The raster is retrieved by the web app from the WMS server and loaded temporarily in Geoserver (part of the Tethys platform).

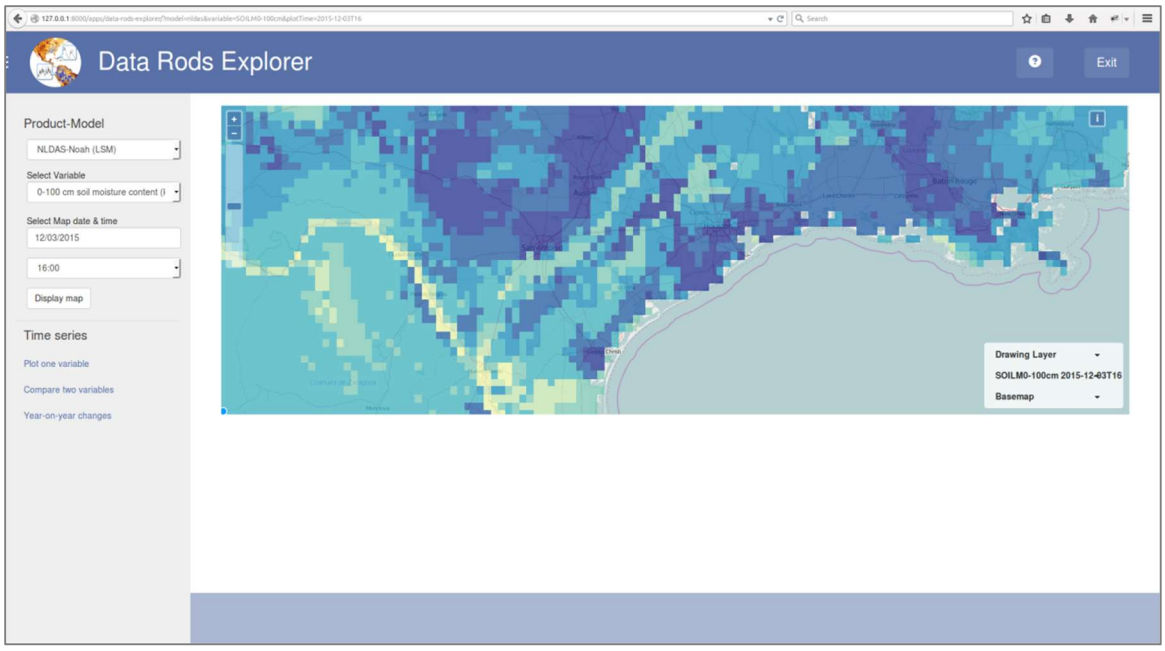


Figure 39: NLDAS-Noah raster for soil moisture in the top meter on December 3, 2015 16:00 UTC. The map is loaded from the WMS based on the GET parameters also populated in the left panel.

Figure 40 shows an example of the interface for the *plot one variable* option on the Data Rods Explorer. The map (top) shows hourly precipitation depth (NLDAS-Noah) in Central Texas on October 31, 2013 01:00 UTC. The plot (bottom) shows the time series of the precipitation depth on the Onion Creek watershed. The time series is retrieved by the web app from the Data Rods server.

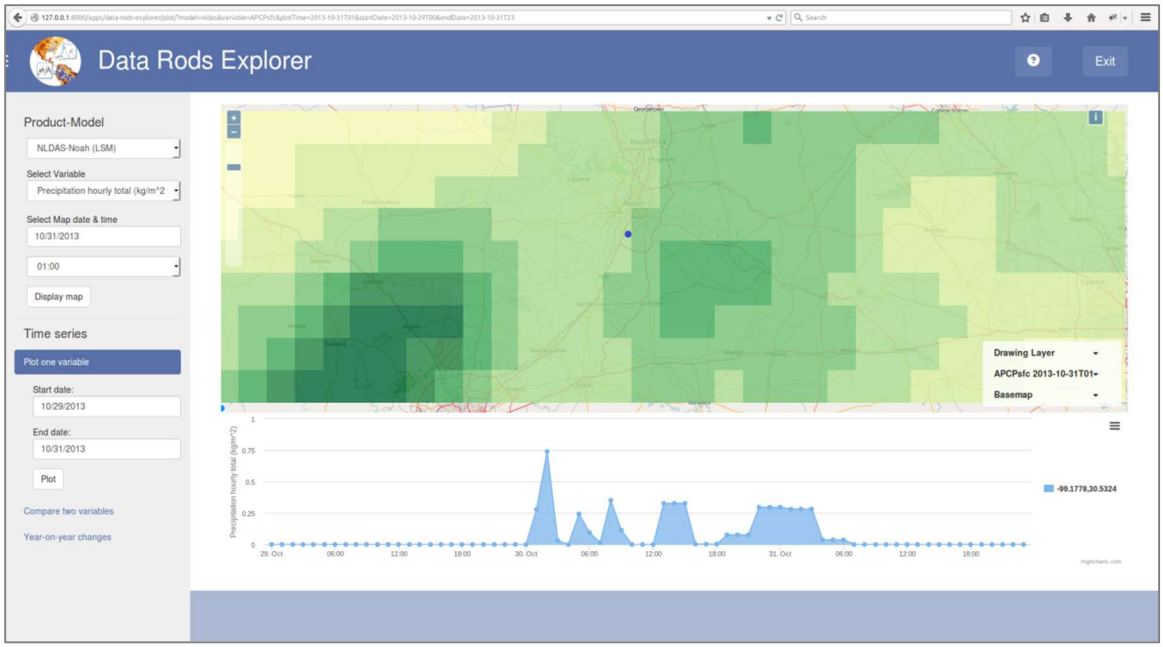


Figure 40: Example of the *plot one variable* option in the Data Rods Explorer. The map shows hourly precipitation during the Halloween flood on October 31, 2013 01:00 UTC on Austin, TX. The plot shows the variation in precipitation depth at the outlet of the Onion Creek watershed on October 29-31, 2013.

Figure 41 shows an example of the *compare two variables* option on the Data Rods Explorer. The plot (bottom) compares surface longwave (blue) and shortwave (black) radiation (NLDAS-Noah) at Tuscaloosa, AL on August 1-15, 2015. The map shows the raster of surface longwave radiation for the southeast on July 1, 2015. As in the previous case, the map and the plot is an integration of data coming from the WMS and the Data Rods servers within the web app.

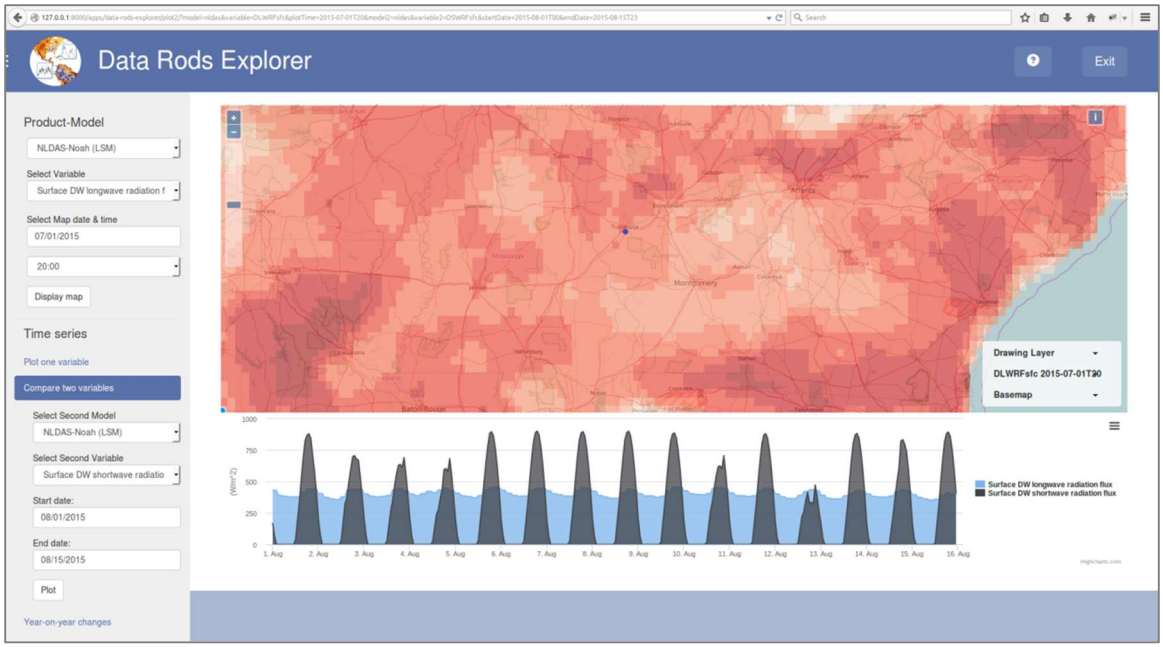


Figure 41: Example of the *compare two variables* option in the Data Rods Explorer. The map shows the surface longwave radiation on Tuscaloosa, AL (blue dot) and the southeast on July 1, 2015. The plot compares the surface longwave and shortwave radiation for Tuscaloosa, AL on August 01-15, 2015.

Figure 42 is an example of the *year-on-year changes* option in the Data Rods Explorer. The plot (bottom) overlays total evapotranspiration data (NLDAS-Noah) at Los Angeles, CA for five years: 2010-2014. The map (top) shows the total evapotranspiration raster for California and the southwest on July 1, 2015. The map shows the persistent low evapotranspiration values (yellow) in comparison from larger values (blue).

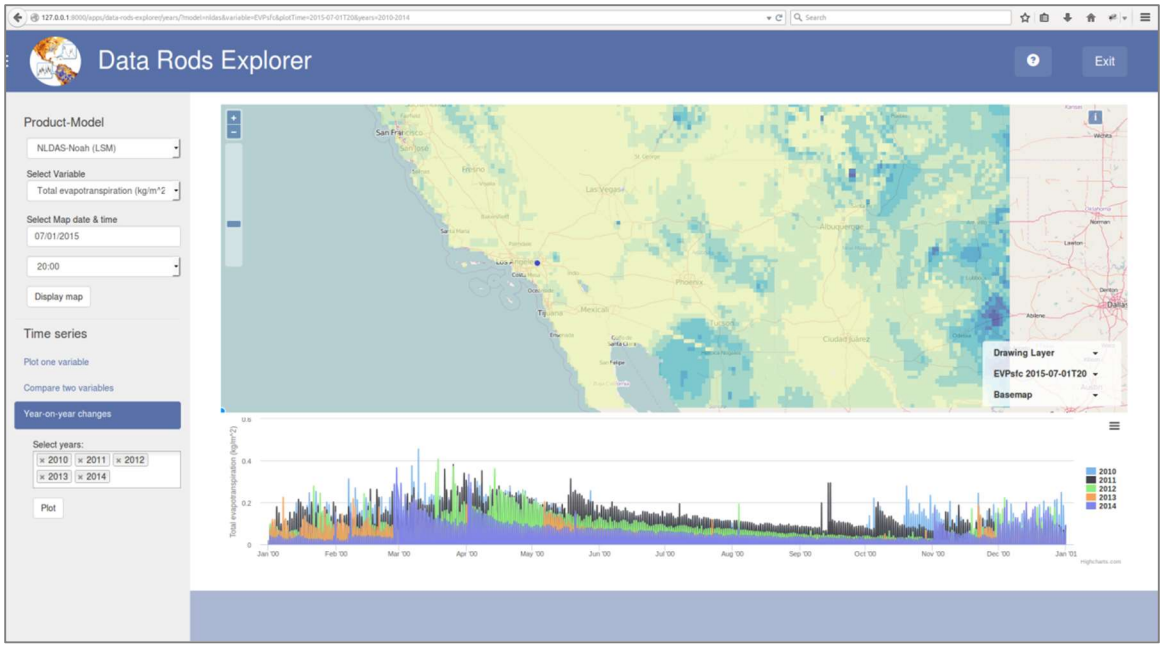


Figure 42: Example of the *year-on-year changes* option on the Data Rods Explorer. The map shows total evapotranspiration in California and the southwest. The plot shows the comparison of total evapotranspiration from 2010 to 2014.

The Data Rods Explorer web app is setup in a way that more models and variables can be added using the same GET parameters structure in the URL. The web app can be easily expanded as more models are available in the Data Rods server. The URLs are structured using the same framework for accessing hydrologic described in Chapter 5. An important feature is that the LDAS models time series are precomputed and the TRMM and GRACE models are computed on-the-fly. The variables computed on-the-fly have a small time lag at plotting the values.

Chapter 5: LDAS Data Access Framework and its Integration in Hydrologic Analysis

The objective of this chapter is to describe thoroughly and present the best practices for data access of NLDAS data using the latest implementation of space-indexed and time-indexed web services. Two application cases are presented: (1) using the GrADS server (space-indexed) to compare current condition in the NLDAS model with the historic trend and (2) the use of the data rods web service (time-indexed) to retrieve input flow data for hydrologic routing.

5.1 LDAS MODELS AND QUERY PARAMETERS

The available LDAS models are identified by their *project name* and *product name*. The *project name* distinguishes between the Global (GLDAS) or North-American (NLDAS) datasets. The *product nam*e depends on the spatial and temporal coverage and resolution.

Table 9 shows a list of the available NLDAS models relevant to the present research, their product names, the spatial and temporal coverage and resolution. The suffix 002 indicates the version of the model (e.g. NLDAS-2). The forcing parameters (FOR, and FORA prefixes) are independent of the model (i.e. Noah, VIC, or Mosaic), hence they are found in a separate product. The *H*, *3H*, and *M* correspond to the temporal resolution (hourly, 3-hourly, and monthly). The combination of the project and product names provides relevant information about the model, the spatial coverage, and the spatial resolution but it did not inform about the spatial resolution and the temporal coverage which have to be queried from the information webpage. A complete list of all product projects and names available in the GrADS server is listed in Appendix I.

Project Name (Spatial Coverage)	Product Name	Temporal Coverage (Start Date)	Spatial Resolution	Temporal Resolution
GLDAS	NOAH10_3H.020	01/01/1948 03Z	1 degree	3-Hourly
GLDAS	NOAH10_M.020	01/01/1984 00Z	1 degree	Monthly
NLDAS	FORA0125_H.002	01/01/1979 13Z	1/8 degree	Hourly
NLDAS	FORA0125_M.002	01/01/1979 00Z	1/8 degree	Monthly
NLDAS	NOAH0125_H.002	01/02/1979 01Z	1/8 degree	Hourly
NLDAS	NOAH0125_M.002	01/01/1979 00Z	1/8 degree	Monthly

Table 9: NLDAS Noah products, their spatial-temporal resolution and coverage (Goddard Space Flight Center, 2015a).

Some of these LDAS datasets and some variables have been implemented as "data rods" (Section 5.2) which are time series for a given point instead, instead of a raster for a given time interval. This means that the data rods server is indexed by time which is suitable for applications that require analyzing the variations of a hydrologic variable in time for a given point in space.

5.2 DATA ACCESS THOUGH TIME-INDEXED WEB SERVICES: DATA RODS

The access of millions of data values and its processing required implementing an optimized methodology for data retrieval. The process was automated using Python scripts which queried the data from the data rods web service, stored it temporarily, and parsed it to an R code that calculated and validated the distributions.

The data was obtained from the NLDAS-2 Noah model for the continental United States for the five variables on a period of 35 years (1979-2013). The data retrieval was made through the data rods web service. The data rods web service provided the time

series for a given cell, meaning that is indexed by time instead of space. In general, the use of data rods improved the data access process for this analysis, because each cell can be processed independently (i.e. in parallel). The data retrieval process was automated and implemented using the High-Performance Computing (HPC) in the Texas Advanced Computer Center (TACC) using the supercomputer Stampede and processing each variable and each cell in parallel.

Figure 43 shows a sample link for accessing LDAS data using the data rods web service (Goddard Earth Sciences Data and Information Services Center, 2015a). The link is structured as a query string were the variable, output format, location, and time extent are specified as GET parameters. The query string structure facilities the automation of the data acquisition process.

```
http://hydro1.sci.gsfc.nasa.gov/daac-bin/access/timeseries.cgi?   -Data Rods server
variable=NLDAS:NLDAS_NOAH0125_H.002:SOILM0-100cm&                 -LDAS Variable
type=asc2&                                                        -Output format
location=GEOM:POINT(-97.6875, 30.1875)&                           -Grid point
startDate=1979-01-02T06&endDate=2014-01-01T05                     -Time extent
```

Figure 43: Example link for accessing LDAS data through the Data Rods web service. The variable (red), output format (green), location (blue), and time extent (purple) are specified by the user.

The variable parameter (Table 10) is constructed by appending the project name (e.g. NLDAS), the product name (e.g. NLDAS_FORA0125.002), and the variable short name (e.g. APCPsfc) separated by a colons. The location can be provided by the index X, Y of the grid or the latitude and longitude (i.e. GEOM:POINT). The date fields are

specified in a simple format (yyyy-mm-ddThh), and the data can be retrieved in four format types: plot, WaterML, netcdf, and ascii.

Key	Value Convention	Example
variable	projectName: productName: VariableShortName	NLDAS:NLDAS_FORA0125.002:APCPsfc
		GLDAS:GLDAS_NOAH025_3H.001:precip
location	projectName:X%3.3d-Y%3.3d	NLDAS:X301-Y080
	projectName:X%4.4d-Y%3.3d	GLDAS:X1220-Y137
	GEOM:POINT(lon, lat)	GEOM:POINT(-100.2, 29.89)
startDate	yyyy-mm-ddThh (default is the 1st time step)	2012-03-30T00
endDate	yyyy-mm-ddThh (default is the last time step)	2012-03-30T23
type	plot	plot (output time series plot)
	asc2	asc2 (output 2-column ASCII)
	asci	ascii (output 4-column ASCII)
	netcdf	netcdf (output netcdf, just a prototype)
	waterml	waterml

Table 10: Key-Value-Pair Syntax (Goddard Earth Sciences Data and Information Services Center, 2015).

Table 11 shows the five variables and their short names (i.e. soil moisture, evapotranspiration, precipitation, runoff, and temperature) used in the multidimensional statistical analysis (Chapter 3). A complete list of the available variables implemented as data rods at the moment of writing is listed in Appendix II.

Data Product	Short Name	Description	Units
NLDAS-2 Primary Forcing	APCPsfc	Precipitation hourly total	kg/m^2
	TMP2m	2-m above ground temperature	K
NLDAS-2 0.125x0.125 Degree Noah LSM Model	EVPsfc	Total evapotranspiration	kg/m^2
	SSRUNsfc	Surface runoff (non-infiltrating)	kg/m^2
	SOILM0-100cm	0-100 cm soil moisture content	kg/m^2

Table 11: NLDAS-2 variables and their access codes (short names) used in the statistical analysis (Goddard Earth Sciences Data and Information Services Center, 2015a).

Figure 44 is an example of the data returned by the link in Figure 43. The ascii file includes all the metadata needed to identify the request and the time series. The time series is be efficiently parsed and the process automated.

```
Metadata of the Time Series file:

prod_name=NLDAS_NOAH0125_H.002
param_short_name=SOILM0-100cm
param_name=0-100 cm top 1 meter soil moisture content
unit=kg/m^2
undef=  9.9990e+20
begin_time=1979/01/02/01
end_time=2015/12/06/00
time_interval[hour]=1
tot_record=323688
grid_y=41 (lat=  30.1875)
grid_x=218 (lon= -97.6875)
elevation[m]=157.600998
dlat=0.125000
dlon=0.125000
ydim(original data set)=224
xdim(original data set)=464
start_lat(original data set)=  25.0625
start_lon(original data set)=-124.9375
Last_update=Thu Dec 10 16:41:19 2015

Metadata for Requested Time Series:

prod_name=NLDAS_NOAH0125_H.002
param_short_name=SOILM0-100cm
param_name=0-100 cm top 1 meter soil moisture content
unit=kg/m^2
begin_time=1979/01/02/06
end_time=2014/01/01/05
begin_time_index=5
end_time_index=306796
lat=  30.1875
lon= -97.6875
grid_y=41
grid_x=218
tot_record=306792
Request_time=Fri Dec 11 14:39:30 2015

           Date&Time       Data
      1979-01-02 06Z     2.4708E+02
      1979-01-02 07Z     2.4708E+02
      1979-01-02 08Z     2.4708E+02
      1979-01-02 09Z     2.4707E+02
      1979-01-02 10Z     2.4707E+02
      1979-01-02 11Z     2.4706E+02
      1979-01-02 12Z     2.4706E+02
         ...                ...
```

Figure 44: Example of the output file from the data rod web service displaying soil moisture data in the top meter

5.3 DATA ACCESS THROUGH SPACE-INDEXED WEB SERVICES: WMS

Geographic data efficiently shares ideas through graphic elements, its power rely on the visualization of a concept in a map. The cost of using geographic data is the relative complexity of the additional information that describes it. Time series data requires only two parameters to be described: time stamp and value. In contrast, geographic data requires certain parameters and rules to be described such as: projection, datum (e.g. WGS84 in LDAS), extent, location, or cell size. This additional complexity is minimized using standards such as Web Mapping Services (WMS) and GeoTIFFs, approved by the Open Geospatial Consortium (OGC).

Figure 45 shows a sample link for accessing LDAS data using the WMS from Giovani (Rui et al., 2011). Similar to the data rods web service, the WMS link is structured as a query string were the parameters are specified as GET variables. The simple and structured link leverages the use of this information within GIS software or web-based applications. The variable parameter is constructed identically as in the data rod service using the parameters in Figure 44 and Figure 45 (or Appendix I and Appendix II.) The image options include the projection (SRS variable) and the width and height of the image. The location is provided by a bounding box (minimum Longitude, minimum Latitude, maximum Longitude, and maximum Latitude). The time parameter is constructed with the start and end date, if the dates are different an image with the average values is returned.

```
http://giovanni.gsfc.nasa.gov/giovanni/daac-bin/wms_ag4?          -WMS server
VERSION=1.1.1&REQUEST=GetMap&SRS=EPSG:4326&WIDTH=512&HEIGHT=256   -Image options
&LAYERS=Time-Averaged.NLDAS_NOAH0125_M_002_soilm0_100cm           -LDAS variable
&STYLES=default&TRANSPARENT=TRUE&FORMAT=image/tiff                -Output format
&time=2008-01-01T00:00:00Z/2008-01-01T00:00:00Z                   -Time extent
&bbox=-119,30,-107,36                                             -Bounding box
```

Figure 45: Example link for accessing LDAS data through the WMS service. The projection and image parameters (orange), the variable (red), output format (green), time extent (purple), and location (blue) are specified by the user.

The format parameter on Figure 45 is the format of the output image file. Table 12 shows the available options for the projection and format parameters at the time of writing (Goddard Earth Sciences Data and Information Services Center, 2015b).

Format	
image/png	image/svg+xml
image/jpeg	image/tiff
image/png; mode=8bit	application/vnd.google-earth.kml+xml
application/x-pdf	application/vnd.google-earth.kmz
SRS (projection)	
EPSG:4326	

Table 12: Output image format and projection in the WMS server (Goddard Earth Sciences Data and Information Services Center, 2015b).

Figure 46 is an example of the data returned by the link in Figure 45. The image file includes the geographic information needed to be displayed correctly in any GIS software or web applications.

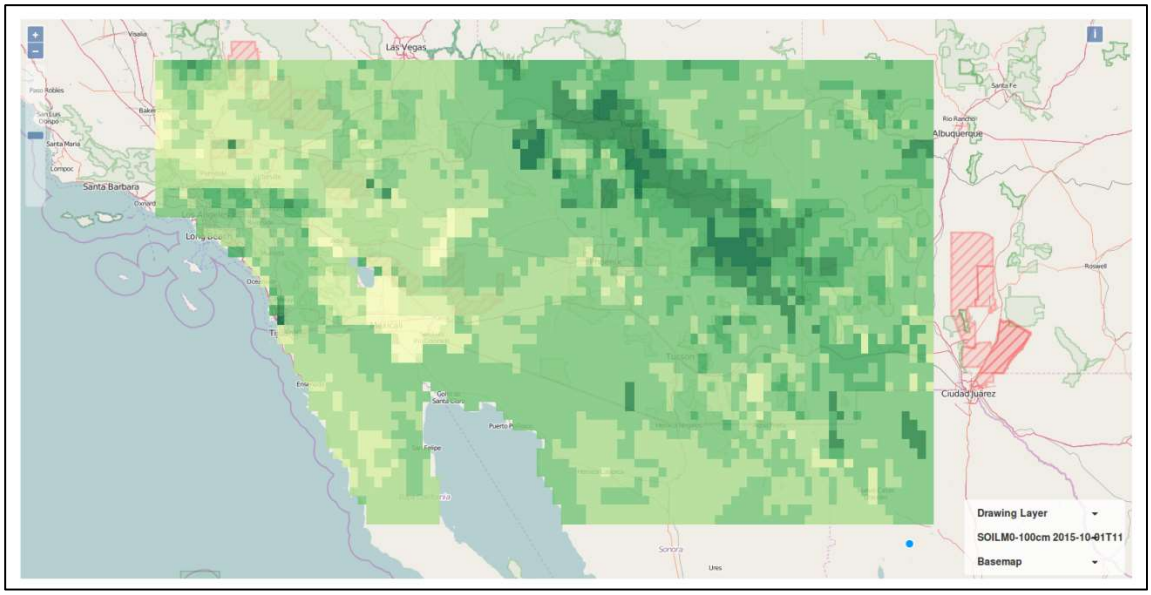


Figure 46: Example image displaying soil moisture in the top meter for south-western United States at 01/01/2008 00:00 UTC.

5.4 RESULTS

The framework presented for retrieving LDAS data through standardized procedures utilizes the best information technologies available in web services and GIS. The strength of the framework relies in: (1) the replication of the data in the time-indexed and the space-indexed servers which improves performance, (2) the setup of the parameters as GET variables in the URL string, (3) the readiness of the data to be integrated with mapping and GIS software, and (4) the capacity to combine the web services and display the data in dynamic web applications. A combination of both approaches (i.e. the time-index and the space-indexed servers) was optimal: querying

time series for a point from the time-indexed server and creating maps for a given time interval from the space-indexed server.

Two application cases were carried out using this framework: (1) Using data services for comparing current conditions with historic values and (2) using data rods as data input for hydraulic routing.

5.4.1 Using data services for comparing current conditions with historic values

The LDAS data access framework was integrated in two websites: (1) Texas soil moisture web app and (2) Stats NLDAS web app. These websites are public and only require an internet connection and a web browser to be accessed. The Texas soil moisture app display is available at http://texassoilmoisture.azurewebsites.net/ (Espinoza, Maidment, García-Martí, & Whiteaker, 2014) and compares the soil moisture values in the top meter of soil in the state of Texas with the historic values. The Stats NLDAS app is available at http://statsnldas.azurewebsites.net/ (Espinoza, Arctur, Maidment, & Teng, 2015) and compares the latest values in five NLDAS variables: (1) soil moisture in the top meter, (2) total evapotranspiration, (3) surface runoff, (4) precipitation, and (5) two meters above ground temperature. The values of the variables are compared to the historic values.

The historic values were obtained with the methodology described in Chapter 3. The web applications were developed and deployed using the methodology described in Chapter 4.

5.4.1.1 Texas soil moisture web app

Figure 47 show the web app of soil moisture values in Texas (Section 4.3.1) for October 23, 2015. The map displays the values above the 80 percentile (blue) and the values below the 20 percentile (red) which is updated when new data is available using

the GrADS web service (this web service was replaced by the WMS server in the NLDAS statistical map). The map is clickable, displaying a pop-up window with the statistical information of the location clicked and updates the two figures on the right. The top-right figure plots the Cumulative Distribution Function (CDF) for the displayed day of the year. The bottom-right figure plots the last 30 day values from the data rods web service and the corresponding 20 and 80 percentiles for the day.

The use of the GrADS server (or the WMS server) eases and automates the process of updating the web application because a tiff raster is retrieved instead of plain text. The tiff raster is queryable, geolocated, and ready-to-use in geoprocessing tools. The use of the data rods web service allows the web app to display and plot the data instantaneously for a point in space. This is because the location is given and the light response (i.e. date time and values) from the server can be parsed and plotted efficiently.

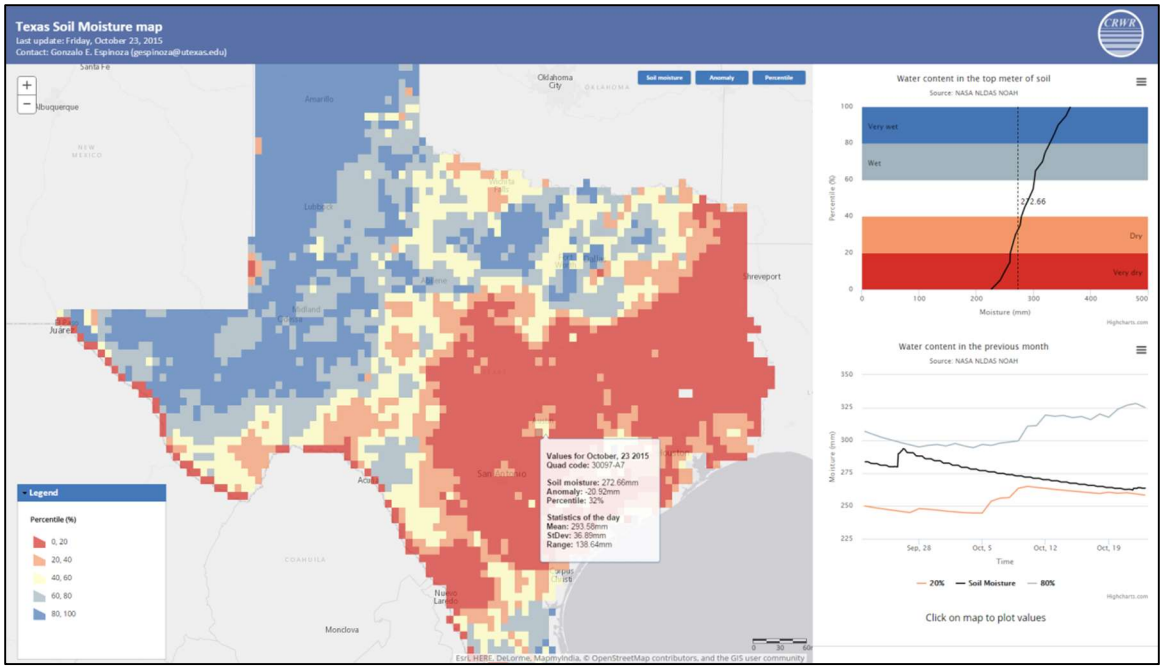


Figure 47: Soil moisture in the top meter percentiles in Texas at October 23, 2015. The web app shows values above the 80 (blue) and below the 20 percentiles (red). The value of location clicked on the map is compared against the historic CDF of the day (top-right) and the previous 30 day values are also compared against the 80 and 20 percentiles (bottom-right).

5.4.1.2 NLDAS statistical webapp

Figure 48 shows the NLDAS statistical web app (section 4.3.2) for October 23, 2015. The map displays the anomaly in soil moisture defined as the number of standard deviations from the daily mean (i.e. standard score) (Equation 12).

$$anom \quad \frac{val - \mu}{\sigma} \qquad (13)$$

Where:

$anom$: is the anomaly as the number of standard deviations from the daily mean.

μ: is the daily mean and σ: is the daily standard deviation.

The regions above the 75 (light-blue) and 90 (dark-blue) percentiles and the regions below the 25 (light-orange) and 10 (dark-orange) percentiles are highlighted. The map quickly displays regions in the continental United States where wetter or drier than usual conditions are occurring. The map includes the ten layers, five for each one of the five variables (soil moisture, evapotranspiration, precipitation, runoff, and temperature) and five for their anomalies. The layers can be changed using the top-right drop-down menus.

The map is clickable, which displays a pop-up with the statistics of the location for the five variables and updates the two figures on the right. Similarly as in the Texas soil moisture web app, the top-right figure plots the current value and compares it to the CDF of the day. The bottom-right figure plots the latest 30 day values for the variable selected on the map. As in the Texas soil moisture app, the integration of the data within the NLDAS statistical app allows the instantaneous mapping and plotting of the data.

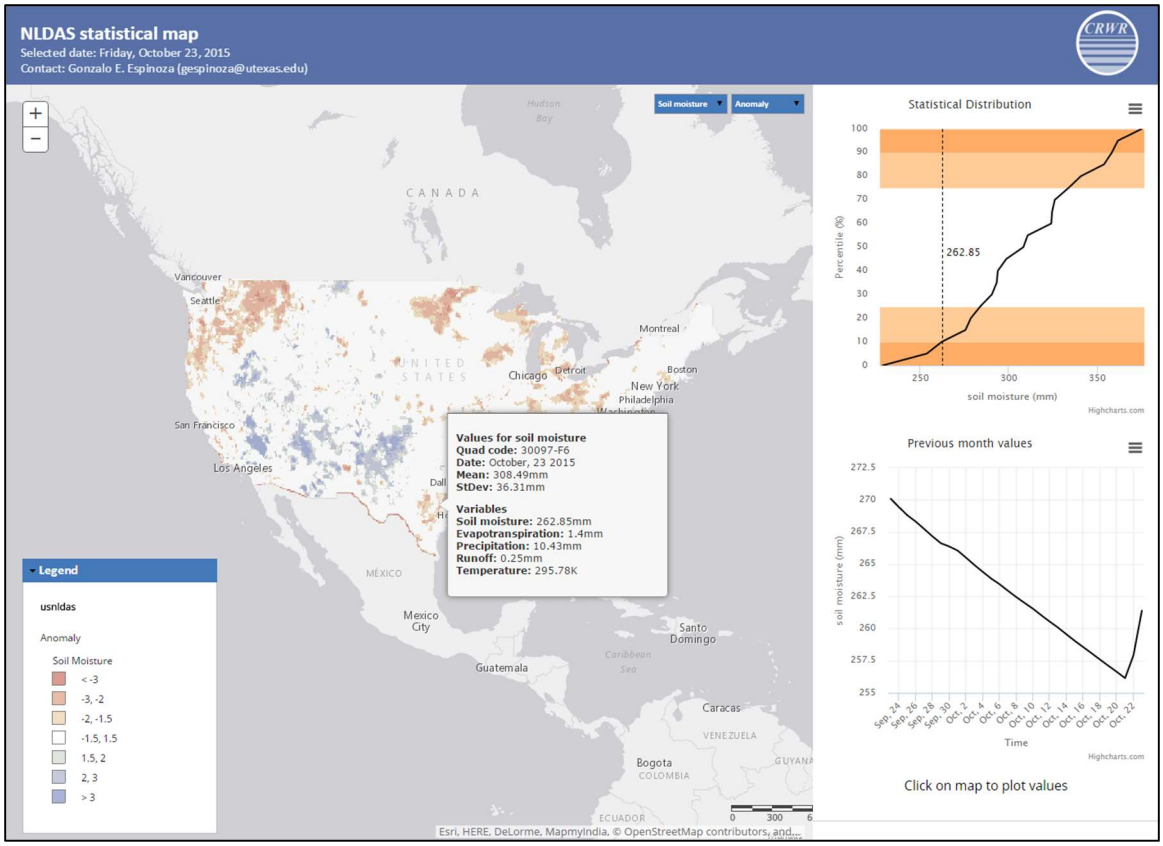


Figure 48: Soil moisture anomaly in number of standard deviations from the daily mean in the continental United States on October, 23 2015. The pop-up on the map displays the values for the five variables at the clicked location. The figure on the top-right shows the soil moisture value and its comparison with the daily CDF. The bottom-right plot shows the previous 30 day values from the data rods web service.

5.4.2 Using data rods as data input for hydraulic routing

NLDAS data had being used to estimate lateral inflow (i.e. surface runoff and baseflow) to river reaches for hydrologic routing applications in: (1) modeling the Upper-Alabama River (Choi et al., 2015) at the National Interoperability Experiment (NFIE), (2) modeling the Onion Creek watershed at the Center for Research in Water Resources (CRWR), and (3) the modeling of the San Antonio-Guadalupe Rivers Basin (Hijar

Santibañez, 2015). The data was obtained using the data rods web service and implemented as a script tool in an ArcGIS toolbox.

The script tool calculates a weighted average of the NLDAS grid cells per drainage area and computes the volume drained to each river reach per hour (m^3/hr). The output inflow file consists in two columns: time stamp and value. The output text file can be transformed into specific inflow files used in hydraulic routing software such as SPRNT (F. Liu & Hodges, 2012) and Rapid (David et al., 2011).

Figure 49 shows the tool dialog on ArcGIS. The input parameters are: (1) the watersheds layer, (2) the field in the watersheds layer containing the COMIDs, (3) the field in the watersheds layer with the drainage areas in square kilometers, (4) the time interval (i.e. start and end dates) of the lateral inflow values, and (5) the NLDAS grid included with the tool. The output parameters are: (1) a table with the areas per river reach, (2) the weights calculated for each drainage area per NLDAS grid cell, and (3) a folder were the output text files that contain the lateral inflow are saved.

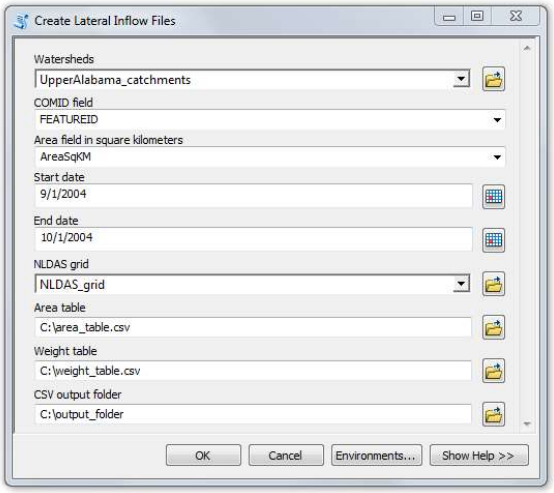


Figure 49: ArcGIS tool dialog for creating the lateral inflow files for a set of drainage areas and time interval. The tool saves creates text files in the output folder and two tables with the areas per river reach and the weights of each NLDAS grid cell for each drainage area.

Figure 50 shows the hydrography of the Upper Alabama River, close to the city of Montgomery, AL. The NLDAS grid (black) is considerable larger than the drainage areas (red) of each river reach (blue) which can induce an additional error that must be accounted by the modelers.

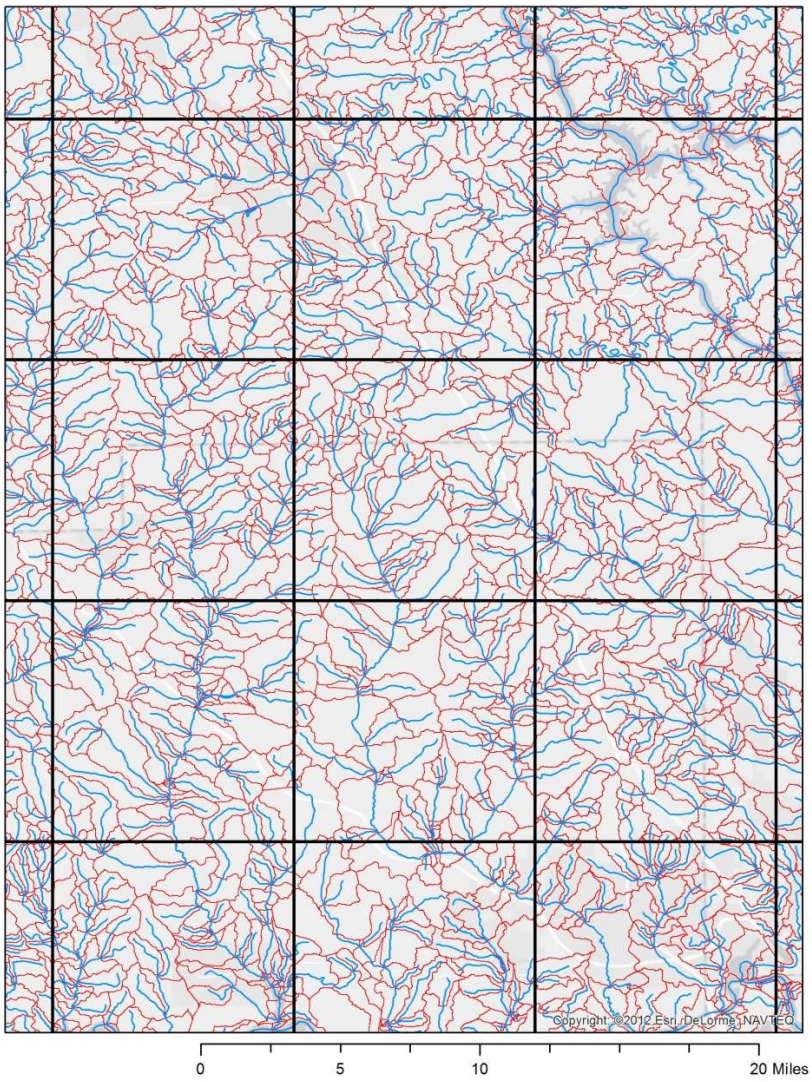


Figure 50: Hydrography of the Upper Alabama River close to the city of Montgomery, AL. The surface runoff and baseflow was obtained from the data rods web service using the NLDAS grid (black) and downscaled to each drainage area (red) of each river reach (blue).

Chapter 6: Conclusions

6.1 Research Summary and Objectives

The research presented in this dissertation provides a novel approach (1) to analyze the spatial-temporal dependence of the probability distributions of hydrologic variables from land surface models, (2) to improve the data exposure of hydrologic information through web applications, and (3) to ease and integrate large hydrologic datasets in practical engineering applications through the use of a data access framework. These novel approaches were applied using data from the North-American Land Data Assimilation System (NLDAS). The research questions, objectives, and scope of this dissertation are:

1. *How can a multi-dimensional analysis of land-surface models improve our understanding of the distribution of hydrologic variables?* A spatial-temporal statistical approach is performed to analyze and interpret the outputs from a land surface hydrologic model (North-American Land Assimilation System NLDAS).

   Objective: Complete a statistical analysis of the NLDAS model output. The statistical analysis includes the summary of the statistics and the calculation of the cumulative distribution functions (CDFs). This is performed for each grid point and for each calendar day and each calendar month, and modeling the CDFs is accomplished using common probability distribution functions. The statistical analysis covers the continental United States, using data from 1979 to 2013 and five variables (soil moisture, evapotranspiration, precipitation, runoff, and temperature) on a 1/8 degree grid with one-day and one-month time steps.

2. *How can hydrologic information be shared dynamically as a final result in an accessible, simple, and interactive approach?* A web application architecture is constructed using the best and latest technologies available; linking web services, cloud deployment and storage, and mapping.

Objective: Create three web map applications for exposing the latest results in NLDAS: (1) latest conditions in soil moisture in Texas and its comparison with the historic trend, (2) statistical map for the continental United States showing the latest conditions and its comparison with historical values for five hydrologic variables (i.e. soil moisture, evapotranspiration, precipitation, runoff, and temperature), and (3) a time series (i.e. data rods) explorer for improving data access and displaying of LDAS data (NLDAS and GLDAS) and two additional global datasets: the Tropical Rainfall Measuring Mission (TRMM) and the Gravity Recovery and Climate Experiment (GRACE).

3. *How can large models datasets can be queried, parsed, and used efficiently in hydrologic analysis?* A detailed web-based process for data retrieval and integration is described and implemented.

Objective: Describe a detailed framework for accessing NLDAS data. This focuses on improving performance depending on the application case using two alternatives: for space-indexed or time-indexed data. Two study cases are carried out: (1) comparing current conditions with long-term historic trend, where space is the main variable and (2) the use of "data rods" (i.e. time series constructed from a given point in space) as input in hydrologic routing, where time is the main variable.

6.2 Objective 1: Statistical Analysis of NLDAS Model Output

The statistical parameters computed (i.e. mean, variance, and percentiles) for each variable, each grid point in NLDAS, and each calendar day and month provide relevant information about usual values and ranges, that can be used in the fields of water management and emergency response. The Cumulative Distribution Functions (CDFs) computed, which are function of the location and the time of the year, are representations of climatologic and seasonal variations.

The use of common families of probability distributions to fit the computed CDFs reduces the data needed to represent the spatial-temporal variability of a hydrologic variable. The fitted distributions are unbounded by the historic minimum and maximum values, hence a probability can be assigned to new or hypothetical events that are outside the historic range of values.

The fitted CDFs proved satisfactory due to the mathematical approach taken, particularly in: (1) allowing the spatial-temporal variability, (2) using different approaches for different hydrologic variables regarding their type: if they are quantities (soil moisture and temperature) or fluxes (evapotranspiration, precipitation and runoff), (3) the selection of the time interval (day and month), (4) the selection of the common distribution family, and (5) a two-step calculation that removes the large amount of zeroes in the calculations for precipitation and runoff. The fitted CDFs smooth out the local variations and fluctuations and provide representative probability distributions of the historic conditions for the five variables.

In the case of the precipitation and runoff variables, the fits were greatly improved using the two-step process (i.e. fitting a Bernoulli distribution if the event of precipitation or runoff occurs and then fitting a probability distribution for the precipitation or runoff depth). A limitation of this methodology (for the precipitation and runoff variables in a

very small proportion of the total number of fits) was found on in dry areas with a small number of days per year with rain. A distribution could not be fitted but it can be inferred from neighboring cells and days.

The fits for evapotranspiration showed in general excellent results but there was a small bias that was identified (for a small proportion of the total number of cells). The evapotranspiration fits for areas and days of the year with low values (close to zero) performed regularly inferior than the rest of the fits. This is due that the gamma distribution forces a positive-only set of values and NLDAS estimate negative values for evapotranspiration in some cases. Even though, the fits obtained using gamma distribution were the ones that performed best in the vast majority of the cases but attention should be put while using it in areas where the NLDAS model output has zero or negative evapotranspiration values.

The computation of the distributions in a daily and a monthly time interval allows capturing hydrologic extremes that occur in different time scales. For example, the monthly CDFs can be used for drought assessment and the daily distributions for flood analysis. These distributions are useful for understanding extreme events and when anomalies are surpassing thresholds

The combination of latest or past results from NLDAS and their computed CDFs allow us to compare hydrologic states with the historic 35-year values and to understand where anomalies are occurring, which can lead to extreme events.

6.3 OBJECTIVE 2: HYDROLOGIC WEB APPLICATIONS

Hydrological sciences can benefit from the field of geographic web applications. Information can be shared and exposed in innovative, seamless, and real-time web applications. The instant mapping and plotting of data are part of the core functionality of the web application architecture presented.

It was learned that the web applications serve as an integration of web services through mapping and plotting. The main advantage is that web sites are dynamic and updated when new data is available. The information displayed and plotted is selected by the users.

The exposure of hydrologic analysis and data in convenient and informative web applications is a backbone for real-time assessment and response of extreme events (e.g. flooding and the NFIE project). The web application architecture facilitates the process of sharing and informing about current hydrologic conditions. Additional effort should be made to bring the Texas and the National web applications to display real-time data instead of latest conditions. The main challenge will be to reduce the lag in NLDAS data that is about a week. Another alternative could be to use forecast information from the National Water Model to estimate current and short-forecast conditions.

The Data Rods Explorer developed using the Tethys platform shows that open-source tools can leverage hydrologic sciences and lower the barrier for web development for water resources engineers. It integrates data from various models (e.g. LDAS, TRMM, and GRACE) and allows comparing the values from the models in the same application (e.g. precipitation from NLDAS versus precipitation from TRMM).

The cloud offers a powerful solution that leverages the implementation and deployment of hydrologic web applications in a simple, scalable, and robust way. It is recommended its use, if the size of a project justifies it.

Both software alternatives analyzed (i.e. the ArcGIS and the Tethys platforms) are suitable for web applications. The selection of either alternative for a given project will depend on the specific characteristics and needs of the project. The Tethys platform is a system with great potential due to the inclusion of mapping, plotting, and geoprocessing components (e.g. PostGIS and Geoserver) in an open-source framework. The challenge in the future for the Tethys platform will be the maintenance of the software and to ease its deployment on applications in the cloud.

6.4 OBJECTIVE 3: NLDAS DATA ACCESS FRAMEWORK

The research shows that the implementation of a common framework for accessing hydrologic data is essential for sharing, displaying, interpreting, and quickly assimilating results from land-surface models into practical engineering applications. It was learned that the double indexation of the data in time and in space facilitates the inclusion of these information in applications.

The NLDAS data can be included in hydrologic web applications using the WMS and the data rods servers. The data does not have to be preloaded and it can be generated automatically depending on the user's query. The servers show great performance and the information can be mapped and plotted instantaneously.

The usability of an extensive model such as NLDAS relies on its accessibility for querying and retrieving information. This was achieved through the standardization of web services and query mechanisms, including the use of GET variables and standard output formats (e.g. WaterML and NetCDF).

It was learned that the framework is adequate to be used engineering applications such as the estimation of lateral inflow in hydraulic routing and the comparison of current and historic conditions using web applications.

6.4 FUTURE WORK

It is recommended to implement the statistical methodology to more datasets (e.g. NLDAS-VIC) and compare the results. It would be proper to compute the CDFs and the fits from an ensemble of models using the methodology presented.

In the NFIE context, further hydrologic and hydraulic analysis can be performed only in the areas with larger precipitation and runoff forecast depths and high soil moisture percentiles. Focusing of computational resources only in pre-identified regions by the statistical analysis can reduce the required computational resources. In the drought monitor context, a thorough comparison and correlation between the identified drought areas and the soil moisture percentiles is recommended.

The statistical analysis provides a vast amount of information in an historical context. Future research will include updating the probability distributions when new data is available, also to estimate the differences in the statistical models for different time intervals or different land-surface models (LSM). A sensitivity analysis might be desirable to estimate how a large event can affect the historic probability distributions. Additional research can study how the CDFs might be affected by climate change. A good way to do this is to include the projections made by the Intergovernmental Panel on Climate Change (IPCC).

The spatial-temporal analysis of the random fields in a watershed can improve (1) the estimates of hydrologic variables such as infiltration rates and surface runoff due differences in soil moisture or (2) the hydrologic modeling of a storm with varying intensity and precipitation depths across the watershed (Vanmarcke, 2010). These two lines of research can be extended based on the statistical analysis presented.

The current hydrologic states of a given point can be linked in time and space to neighboring cells. An estimation of the relationship, strength, and influence of the neighbors can be used to compare if the conditions for a given hydrologic variable are more related in time or in space.

The TRMM and GRACE variables in the Data Rods Explorer web app are created on-the-fly instead of being pre-computed in the Data Rods server. This process generates some time lag while plotting the variables and diminished the app performance. Additional work would considerer the pre-computation of these datasets.

The WMS server returns the rasters queried with some time lag. This lag can be perceived by the users as a deficient implementation in the web applications. The WMS server could return rasters in almost real-time to avoid this issue.

It is recommended to expand the data rods server to include more models and variables. Especially forecast data from the NFIE project, which could show what areas of the United States, might be subject of flooding in the upcoming days.

Appendix I: LDAS dataset products

Project Name (Spatial Coverage)	Product Name	Temporal Coverage (Start Date)	Spatial Resolution	Temporal Resolution
GLDAS	MOS10SUBP_3H	01/02/1979 00Z	1 degree	3-Hourly
GLDAS	MOS10_M	01/01/1979 00Z	1 degree	Monthly
GLDAS	NOAH025SUBP_3H	02/24/2000 00Z	1/4 degree	3-Hourly
GLDAS	NOAH025_3H.020	01/01/1948 03Z	1/4 degree	3-Hourly
GLDAS	NOAH025_M	03/01/2000 00Z	1/4 degree	Monthly
GLDAS	NOAH025_M.020	01/01/1948 00Z	1/4 degree	Monthly
GLDAS	NOAH10SUBP_3H	01/02/1979 00Z	1 degree	3-Hourly
GLDAS	NOAH10_3H.020	01/01/1948 03Z	1 degree	3-Hourly
GLDAS	NOAH10_M	01/01/1979 00Z	1 degree	Monthly
GLDAS	NOAH10_M.020	01/01/1984 00Z	1 degree	Monthly
GLDAS	VIC10_3H	01/01/1979 03Z	1 degree	3-Hourly
GLDAS	VIC10_M	01/01/1979 00Z	1 degree	Monthly
NLDAS	FOR0125_H.001	08/01/1996 00Z	1/8 degree	Hourly
NLDAS	FOR0125_M.001	08/01/1996 00Z	1/8 degree	Monthly
NLDAS	FOR0125_MC.001	01/01/1997 00Z	1/8 degree	Monthly
NLDAS	FORA0125_H.002	01/01/1979 13Z	1/8 degree	Hourly
NLDAS	FORA0125_M.002	01/01/1979 00Z	1/8 degree	Monthly
NLDAS	FORA0125_MC.002	01/01/1980 00Z	1/8 degree	Monthly
NLDAS	FORB0125_H.002	01/01/1979 13Z	1/8 degree	Hourly

NLDAS	FORB0125_M.002	01/01/1979 00Z	1/8 degree	Monthly
NLDAS	FORB0125_MC.002	01/01/1980 00Z	1/8 degree	Monthly
NLDAS	MOS0125_H.002	01/02/1979 00Z	1/8 degree	Hourly
NLDAS	MOS0125_M.002	01/01/1979 00Z	1/8 degree	Monthly
NLDAS	MOS0125_MC.002	01/01/1980 00Z	1/8 degree	Monthly
NLDAS	NOAH0125_H.002	01/02/1979 01Z	1/8 degree	Hourly
NLDAS	NOAH0125_M.002	01/01/1979 00Z	1/8 degree	Monthly
NLDAS	NOAH0125_MC.002	01/01/1980 00Z	1/8 degree	Monthly
NLDAS	VIC0125_H.002	01/02/1979 00Z	1/8 degree	Hourly
NLDAS	VIC0125_M.002	01/01/1979 00Z	1/8 degree	Monthly
NLDAS	VIC0125_MC.002	01/01/1980 00Z	1/8 degree	Monthly

Table 13: Complete list of LDAS products and their spatial-temporal resolution. The prefixes: NOAH, VIC, and MOS refer to the Noah, VIC, and Mosaic models respectively. FORA/B are the forcing parameters (Goddard Space Flight Center, 2015b).

Appendix II: Variables Available as Data Rods

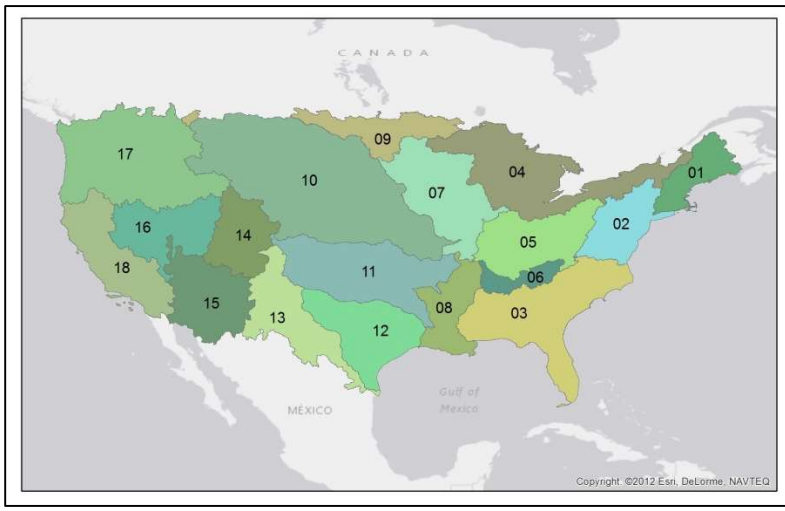
Data Product	Short Name	Description	Unit
NLDAS-2 Primary Forcing	APCPsfc	Precipitation hourly total	kg/m^2
	TMP2m	2-m above ground temperature	K
	DLWRFsfc	Surface DW longwave radiation flux	W/m^2
	DSWRFsfc	Surface DW shortwave radiation flux	W/m^2
	PEVAPsfc	Potential evaporation	kg/m^2
	SPFH2m	2-m above ground specific humidity	kg/kg
	TMP2m	2-m above ground temperature	K
	UGRD10m	10-m above ground zonal wind	m/s
	VGRD10m	10-m above ground meridional wind	m/s
NLDAS-2 0.125x0.125 Degree Noah LSM Model	EVPsfc	Total evapotranspiration	kg/m^2
	GFLUXsfc	Ground heat flux	w/m^2
	LHTFLsfc	Latent heat flux	w/m^2
	SHTFLsfc	Sensible heat flux	w/m^2
	SSRUNsfc	Surface runoff (non-infiltrating)	kg/m^2
	BGRIUNdfc	Subsurface runoff (baseflow)	kg/m^2
	SOILM0-10cm	0-10 cm soil moisture content	kg/m^2
	SOILM0-100cm	0-100 cm soil moisture content	kg/m^2
	SOILM0-200cm	0-200 cm soil moisture content	kg/m^2
	SOILM10-40cm	10-40 cm soil moisture content	kg/m^2
	SOILM40-100cm	40-100 cm soil moisture content	kg/m^2

	SOILM100-200cm	100-200 cm soil moisture content	kg/m^2
	TSOIL0-10cm	0-10 cm soil temperature	K
GLDAS-1 0.25x0.25 Degree Noah LMS Model	Evap	Total Evapotranspiration	kg/m^2/s
	precip	Precipitation rate	kg/m^s/hr
	Rainf	Rain rate	kg/m^2/s
	Snowf	Snow rate	kg/m^2/s
	Qs	Surface Runoff	kg/m^2/s
	Qsb	Subsurface Runoff	kg/m^2/s
	SOILM0-100cm	0-100 cm top 1 meter soil moisture content	kg/m^2
	SOILM0-10cm	0-10 cm layer 1 soil moisture content	kg/m^2
	SOILM10-40cm	10-40 cm layer 2 soil moisture content	kg/m^2
	SOILM40-100cm	40-100 cm layer 3 soil moisture content	kg/m^2
	Tair	Near surface air temperature	K
	TSOIL0-10cm	Average layer 1 soil temperature	K
	Wind	Near surface wind magnitude	m/s

Table 14: Complete list of variables recognized as time series (Goddard Earth Sciences Data and Information Services Center, 2015a).

Appendix III: Hydrologic Regions in the United States

Figure 51 shows the map and a table with the codes and names of the hydrologic regions in the continental United States from the National Hydrography Dataset (Simley & Carswell Jr., 2009).



HUC2	NAME
07	Upper Mississippi Region
09	Souris-Red-Rainy Region
18	California Region
14	Upper Colorado Region
01	New England Region
15	Lower Colorado Region
02	Mid Atlantic Region
08	Lower Mississippi Region
11	Arkansas-White-Red Region
03	South Atlantic-Gulf Region
06	Tennessee Region
13	Rio Grande Region
16	Great Basin Region
12	Texas-Gulf Region
04	Great Lakes Region
17	Pacific Northwest Region
05	Ohio Region
10	Missouri Region

Figure 51: Hydrologic Regions in the National Hydrography Dataset (NHD).

Appendix IV: Data Rods explorer models, variables, and access codes

Model	GET code	Variable	GET code
NLDAS-Noah	nldas	Precipitation hourly total (kg/m^2)	APCPsfc
		Surface DW longwave radiation flux (W/m^2)	DLWRFsfc
		Surface DW shortwave radiation flux (W/m^2)	DSWRFsfc
		Potential evaporation (kg/m^2)	PEVAPsfc
		2-m above ground specific humidity (kg/kg)	SPFH2m
		2-m above ground temperature (K)	TMP2m
		10-m above ground zonal wind (m/s)	UGRD10m
		10-m above ground meridional wind (m/s)	VGRD10m
		Total evapotranspiration (kg/m^2)	EVPsfc
		Ground heat flux (w/m^2)	GFLUXsfc
		Latent heat flux (w/m^2)	LHTFLsfc
		Sensible heat flux (w/m^2)	SHTFLsfc
		Surface runoff (non-infiltrating) (kg/m^2)	SSRUNsfc
		Subsurface runoff (baseflow) (kg/m^2)	BGRIUNdfc

		0-10 cm soil moisture content (kg/m^2)	SOILM0-10cm
		0-100 cm soil moisture content (kg/m^2)	SOILM0-100cm
		0-200 cm soil moisture content (kg/m^2)	SOILM0-200cm
		10-40 cm soil moisture content (kg/m^2)	SOILM10-40cm
		40-100 cm soil moisture content (kg/m^2)	SOILM40-100cm
		100-200 cm soil moisture content (kg/m^2)	SOILM100-200cm
		0-10 cm soil temperature (K)	TSOIL0-10cm
GLDAS-Noah	gldas	Total Evapotranspiration (kg/m^2/s)	Evap
		Precipitation rate (kg/m^s/hr)	precip
		Rain rate (kg/m^2/s)	Rainf
		Snow rate (kg/m^2/s)	Snowf
		Surface Runoff (kg/m^2/s)	Qs
		Subsurface Runoff (kg/m^2/s)	Qsb
		0-100 cm top 1 meter soil moisture content (kg/m^2)	SOILM0-100cm
		0-10 cm layer 1 soil moisture content (kg/m^2)	SOILM0-10cm

		10-40 cm layer 2 soil moisture content (kg/m^2)	SOILM10-40cm
		40-100 cm layer 3 soil moisture content (kg/m^2)	SOILM40-100cm
		Near surface air temperature (K)	Tair
		Average layer 1 soil temperature (K)	TSOIL0-10cm
		Near surface wind magnitude (m/s)	Wind
TRMM	trmm	Precipitation (mm/hr)	precip
GRACE	grace	Surface Soil Moisture Percentile	surf
		Root Zone Soil Moisture Percentile	root
		Ground Water Percentile	deep

Table 15: Available models and variables in the Data Rods Explorer

References

Alameh, N. (2003). Chaining Geographic Information Web Services. *IEEE Internet Computing*, *7*(5), 22–29. http://doi.org/10.1109/MIC.2003.1232514

ArcGIS Resources. (2015). Key concepts for sharing an image service. Retrieved December 1, 2015, from http://resources.arcgis.com/EN/HELP/MAIN/10.2/index.html#//009t000000p000000

Australian Goverment - Bureau of Meteorology. (2016). Australian Landscape Water Balance. Retrieved June 20, 2001, from http://www.bom.gov.au/water/landscape/

Beguería, S. (2005). Uncertainties in partial duration series modelling of extremes related to the choice of the threshold value. *Journal of Hydrology*, *303*(1-4), 215–230. http://doi.org/10.1016/j.jhydrol.2004.07.015

Beniston, M., Stoffel, M., Harding, R., Kernan, M., Ludwig, R., Moors, E., … Tockner, K. (2012). Obstacles to data access for research related to climate and water: Implications for science and EU policy-making. *Environmental Science & Policy*, *17*, 41–48. http://doi.org/10.1016/j.envsci.2011.12.002

Berman, F., Chien, A., Cooper, K., Dongarra, J., Foster, I., Gannon, D., … Wolski, R. (2001). The GrADS Project: Software Support for High-Level Grid Application Development. *International Journal of High Performance Computing Applications*, *15*(4), 327–344. http://doi.org/10.1177/109434200101500401

Blankinship, J. C., Meadows, M. W., Lucas, R. G., & Hart, S. C. (2014). Snowmelt timing alters shallow but not deep soil moisture in the Sierra Nevada. *Water Resources Research*, *50*(2), 1448–1456. http://doi.org/10.1002/2013WR014541

Botts, M., Percivall, G., Reed, C., & Davidson, J. (2008). OGC® Sensor Web Enablement: Overview and High Level Architecture. In S. Nittel, A. Labrinidis, & A. Stefanidis (Eds.), *GeoSensor Networks SE - 10* (Vol. 4540, pp. 175–190). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-79996-2_10

Choi, C. C., Shang, P., Sung, K., Vimal, S., Yu, C.-W., & Zheng, X. (2015). River Geometry Information Interoperability for Large Scale Hydraulic Modelling. In *3rd CUAHSI Conference on HydroInformatics*. Tuscaloosa, AL.

Coles, S., Pericchi, L. R., & Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, *273*(1-4), 35–50.

http://doi.org/10.1016/S0022-1694(02)00353-0

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises Tests. *The Annals of Mathematical Statistics*, *28*(4), 823–838 CR – Copyright © 1957 Institute of Ma. http://doi.org/10.2307/2237048

David, C. H., Maidment, D. R., Niu, G.-Y., Yang, Z.-L., Habets, F., & Eijkhout, V. (2011). River Network Routing on the NHDPlus Dataset. *Journal of Hydrometeorology*, *12*(5), 913–934. http://doi.org/10.1175/2011JHM1345.1

de la Beaujardiere, J. (2006). *OpenGIS® Web Map Server Implementation Specification*. *Open Geospatial Consortium Inc.* Retrieved from http://www.opengeospatial.org/standards/wms

Dragićević, S. (2004). The potential of Web-based GIS. *Journal of Geographical Systems*, *6*(2), 79–81. http://doi.org/10.1007/s10109-004-0133-4

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., … Tarpley, J. D. (2003). Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research: Atmospheres*, *108*(D22), n/a–n/a. http://doi.org/10.1029/2002JD003296

El Adlouni, S., Bobée, B., & Ouarda, T. B. M. J. (2008). On the tails of extreme event distributions in hydrology. *Journal of Hydrology*, *355*(1-4), 16–33. http://doi.org/10.1016/j.jhydrol.2008.02.011

Entekhabi, D., & Rodriguez-Iturbe, I. (1994). Analytical framework for the characterization of the space-time variability of soil moisture. *Advances in Water Resources*, *17*(1-2), 35–45. http://doi.org/10.1016/0309-1708(94)90022-1

Espinoza, G., Arctur, D., Maidment, D., & Teng, W. (2015). Exposing Probabilistic Hydrologic Information through A Dynamic Web Application using NLDAS data. In *CyberGIS All hands meeting*. Reston, VA.

Espinoza, G., Arctur, D., Teng, W., Maidment, D., García-Martí, I., & Comair, G. (2015). Studying Soil Moisture at a National Level through Statistical Analysis of NASA NLDAS Data. *Journal of Hydroinformatics*.

Espinoza, G., Maidment, D., García-Martí, I., & Whiteaker, T. (2014). A Statistical Analysis and Geospatial Distribution of Soil Moisture in Texas. In *2nd International Conference on CyberGIS*. Redlands, CA.

Esri. (2015). ArcGIS API for Javascript. Retrieved from https://developers.arcgis.com/javascript/jshelp/

Funk, C., Michaelsen, J., Verdin, J., Artan, G., Husak, G., Senay, G., … Magadazire, T. (2003). The collaborative historical African rainfall model: description and evaluation. *International Journal of Climatology*, *23*(1), 47–66. http://doi.org/10.1002/joc.866

Global Precipitation Climatology Centre. (2016). Global Precipitation Climatology Centre. Retrieved June 20, 2001, from https://www.dwd.de/EN/ourservices/gpcc/gpcc.html

Goddard Earth Sciences Data and Information Services Center. (2015a). Data Rods (Time Series Data). Retrieved November 1, 2015, from http://disc.sci.gsfc.nasa.gov/hydrology/data-rods-time-series-data

Goddard Earth Sciences Data and Information Services Center. (2015b). OGC WMS for NASA Giovanni. Retrieved November 1, 2015, from http://giovanni.gsfc.nasa.gov/giovanni/daac-bin/wms_ag4?VERSION=1.1.1&REQUEST=Getcapabilities&service=wms

Goddard Space Flight Center. (2015a). GES DISC GrADS Data Server - GLDAS and NLDAS products. Retrieved November 1, 2015, from http://agdisc.gsfc.nasa.gov/dods/

Goddard Space Flight Center. (2015b). GES DISC GrADS Data Server - GLDAS and NLDAS products.

Groisman, P., Karl, T., Easterling, D., Knight, R., Jamason, P., Hennessy, K., … Zhai, P.-M. (1999). Changes in the Probability of Heavy Precipitation: Important Indicators of Climatic Change. *Climatic Change*, *42*(1), 243–283. http://doi.org/10.1023/A:1005432803188

Highcharts developing team. (2015). Highcharts: Make your data come alive. Retrieved from http://www.highcharts.com/

Hijar Santibañez, A. R. (2015). *Evaluating river cross section geometry for a hydraulic river routing model : Guadalupe and San Antonio river basins*. The University of Texas at Austin.

Husak, G. J., Michaelsen, J., & Funk, C. (2007). Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications.

International Journal of Climatology, *27*(7), 935–944. http://doi.org/10.1002/joc.1441

Jiang, P., Gautam, M. R., Zhu, J., & Yu, Z. (2013). How well do the GCMs/RCMs capture the multi-scale temporal variability of precipitation in the Southwestern United States? *Journal of Hydrology*, *479*, 75–85. http://doi.org/10.1016/j.jhydrol.2012.11.041

Jones, N., Nelson, J., Swain, N., Christensen, S., Tarboton, D., & Dash, P. (2014). Tethys: A Software Framework for Web-Based Modeling and Decision Support Applications. In *International Environmental Modelling and Software Societry (iEMSs)*. San Diego, CA.

Katz, R. W., Parlange, M. B., & Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, *25*(8-12), 1287–1304. http://doi.org/10.1016/S0309-1708(02)00056-8

Lakshmi, V. (2004). The role of satellite remote sensing in the Prediction of Ungauged Basins. *Hydrological Processes*, *18*(5), 1029–1034. http://doi.org/10.1002/hyp.5520

Lakshmi, V., Piechota, T., Narayan, U., & Tang, C. (2004). Soil moisture as an indicator of weather extremes. *Geophysical Research Letters*, *31*(11), n/a–n/a. http://doi.org/10.1029/2004GL019930

Lautenbacher, C. C. (2006). The Global Earth Observation System of Systems: Science Serving Society. *Space Policy*, *22*(1), 8–11. http://doi.org/http://dx.doi.org/10.1016/j.spacepol.2005.12.004

Liu, F., & Hodges, B. R. (2012). Dynamic river network simulation at large scale. *Proceedings of the 49th Annual Design Automation Conference on - DAC '12*, 723. http://doi.org/10.1145/2228360.2228491

Liu, Y., Padmanabhan, A., & Wang, S. (2015). CyberGIS Gateway for enabling data-rich geospatial research and education. *Concurrency and Computation: Practice and Experience*, *27*(2), 395–407. http://doi.org/10.1002/cpe.3256

Maidment, D. (2015). *A Conceptual Framework for the National Flood Interoperability Experiment*. Austin, TX. Retrieved from https://www.cuahsi.org/Files/Pages/documents/13623/nfieconceptualframework_revised_feb_9.pdf

Maidment, D. R. (1993). *Handbook of hydrology*. New York: McGraw-Hill.

Marsaglia, G., Tsang, W. W., & Wang, J. (2003). Evaluating Kolmogorov's Distribution. *Journal of Statistical Software*, *8*(18), 1–4. Retrieved from http://www.jstatsoft.org/v08/i18

Michaelis, C., & Ames, D. (2009). Evaluation and Implementation of the OGC Web Processing Service for Use in Client-Side GIS. *GeoInformatica*, *13*(1), 109–120. http://doi.org/10.1007/s10707-008-0048-1

Miskus, D., NDMC, USDA, & NOAA. (2015). U.S. Drought Monitor. Retrieved from http://droughtmonitor.unl.edu/

Mitchell, K. E. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research*, *109*(D7), D07S90. http://doi.org/10.1029/2003JD003823

Moody, E. G., King, M. D., Schaaf, C. B., & Platnick, S. (2008). MODIS-Derived Spatially Complete Surface Albedo Products: Spatial and Temporal Pixel Distribution and Zonal Averages. *Journal of Applied Meteorology and Climatology*, *47*(11), 2879–2894. http://doi.org/10.1175/2008JAMC1795.1

New, M., Hulme, M., & Jones, P. (1999). Representing Twentieth-Century Space–Time Climate Variability. Part I: Development of a 1961–90 Mean Monthly Terrestrial Climatology. *Journal of Climate*, *12*(3), 829–856. http://doi.org/10.1175/1520-0442(1999)012<0829:RTCSTC>2.0.CO;2

New, M., Hulme, M., & Jones, P. (2000). Representing Twentieth-Century Space–Time Climate Variability. Part II: Development of 1901–96 Monthly Grids of Terrestrial Surface Climate. *Journal of Climate*, *13*(13), 2217–2238. http://doi.org/10.1175/1520-0442(2000)013<2217:RTCSTC>2.0.CO;2

Open Geospatial Consortium. (2014). About OGC. Retrieved from http://www.opengeospatial.org/ogc

OpenLayers development team. (2015). OpenLayers 3: A high-performance, feature-packed library for all your mapping needs. Retrieved May 20, 2012, from http://openlayers.org/

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.r-project.org/

Rajkumar, R. (Raj), Lee, I., Sha, L., & Stankovic, J. (2010). Cyber-physical Systems: The

Next Computing Revolution. In *Proceedings of the 47th Design Automation Conference* (pp. 731–736). New York, NY, USA: ACM. http://doi.org/10.1145/1837274.1837461

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., … Toll, D. (2004). The Global Land Data Assimilation System. *Bulletin of the American Meteorological Society*, *85*(3), 381–394. http://doi.org/10.1175/BAMS-85-3-381

Rodell, M., Mocko, D., & Beaudoing, H. K. (2015). LDAS: Land Data Assimilation Systems. Retrieved May 20, 2011, from http://ldas.gsfc.nasa.gov/index.php

Rui, H., Strub, R., Teng, W., Vollmer, B., Mocko, D., Maidment, D., & Whiteaker, T. (2013). *Enhancing Access to and Use of NASA Hydrological Data*. *AGU Fall Meeting*. Retrieved from http://disc.sci.gsfc.nasa.gov/additional/publications/year_2013/access_hydrological_data.pdf

Rui, H., Teng, W., Vollmer, B., Mocko, D., Beaudoing, H. K., & Rodell, M. (2011). NASA Giovanni Portals for NLDAS/GLDAS Online Visualization, Analysis, and Intercomparison. In *AGU Fall Meeting*. San Francisco, CA. Retrieved from http://disc.sci.gsfc.nasa.gov/additional/publications/giovanni_nldas_gldas_intercomparison.pdf

Sager, T. W. (2010). Kolmogorov-Smirnov Test. Encyclopedia of Research Design. SAGE Publications, Inc. SAGE Publications, Inc. http://doi.org/http://dx.doi.org/10.4135/9781412961288

Simley, J. D., & Carswell Jr., W. J. (2009). *The National Map - Hydrography: U.S. Geological Survey Fact Sheet 2009-3054, 4 p.*

Simpson, J., Kummerow, C., Tao, W.-K., & Adler, R. F. (1996). On the Tropical Rainfall Measuring Mission (TRMM). *Meteorology and Atmospheric Physics*, *60*(1-3), 19–36. http://doi.org/10.1007/BF01029783

Stollberg, B., & Zipf, A. (2007). OGC Web Processing Service Interface for Web Service Orchestration Aggregating Geo-processing Services in a Bomb Threat Scenario. In J. M. Ware & G. Taylor (Eds.), *Web and Wireless Geographical Information Systems SE - 18* (Vol. 4857, pp. 239–251). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-76925-5_18

Tapley, B. D., Bettadpur, S., Watkins, M., & Reigber, C. (2004). The gravity recovery

and climate experiment: Mission overview and early results. *Geophysical Research Letters*, *31*(9), 1–4. http://doi.org/10.1029/2004GL019920

Texas Advanced Computing Center, T., & The University of Texas at Austin, U. (2015). Texas Advanced Computing Center. Retrieved from https://portal.tacc.utexas.edu/tacc-citation

Valentine, D., Taylor, P., & Zaslavsky, I. (2012). WaterML, an Information Standard for the Exchange of in-situ hydrological observations. In A. Abbasi & N. Giesen (Eds.), *EGU General Assembly Conference Abstracts* (Vol. 14, p. 13275).

Vanmarcke, E. (2010). *Random Fields: Analysis and Synthesis*. World Scientific. Retrieved from https://books.google.com/books?id=0MCxDV1bonAC

Vaze, J., Viney, N., Stenson, M., Renzullo, L., van Dijk, A., Dutta, D., … Daamen, C. (2013). The Australian Water Resource Assessment Modelling System (AWRA). *20th International Congress on Modelling and Simulation*, (December 2013), 2506–2512. Retrieved from www.mssanz.org.au/modsim2013/L17/vaze.pdf\nhttps://www.researchgate.net/publication/259182250_The_Australian_Water_Resource_Assessment_Modelling_System_%28AWRA%29

Viglione, A., Chirico, G. B., Komma, J., Woods, R., Borga, M., & Blöschl, G. (2010). Quantifying space-time dynamics of flood event types. *Journal of Hydrology*, *394*(1–2), 213–229. http://doi.org/http://dx.doi.org/10.1016/j.jhydrol.2010.05.041

Wang, S., McKenney, D. W., Shang, J., & Li, J. (2014). A national-scale assessment of long-term water budget closures for Canada's watersheds. *Journal of Geophysical Research: Atmospheres*, n/a–n/a. http://doi.org/10.1002/2014JD021951

Wigley, T. M. L. (2009). The effect of changing climate on the frequency of absolute extreme events. *Climatic Change*, *97*(1-2), 67–76. http://doi.org/10.1007/s10584-009-9654-7

Xia, Y., Ek, M., Wei, H., & Meng, J. (2012). Comparative analysis of relationships between NLDAS-2 forcings and model outputs. *Hydrological Processes*, *26*(3), 467–474. http://doi.org/10.1002/hyp.8240

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., … Lohmann, D. (2012). Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2.

Validation of model-simulated streamflow. *Journal of Geophysical Research*, *117*, D03110. http://doi.org/10.1029/2011JD016051

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., … Mocko, D. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research*, *117*(D3), D03109. http://doi.org/10.1029/2011JD016048

Vita

Gonzalo E. Espinoza Dávalos was born in San Luis Potosi, Mexico in 1986 although earlier in his life he moved to Aguascalientes; the city he considers his hometown. Gonzalo is the middle son of Berenice Dávalos Estrada and Gonzalo Espinoza Del Río. Gonzalo's elder sister is Miriam and his younger brother is Antonio. Gonzalo's maternal family is from the region of Los Altos in the state of Jalisco, and his paternal family is from Zamora in the state of Michoacán.

Gonzalo attended the Autonomous University of Aguascalientes where he received his bachelor's degree in Civil Engineering in June, 2009. In August, 2010 he moved to the United States to attend graduate school at the University of Texas at Austin. He received his Master's degree in Environmental and Water Resources Engineering in May, 2012 and moved forward to pursue a Ph.D. degree in Civil Engineering.

Gonzalo would like to focus his career on improving and solving problems related to water management in relegated regions and communities. He is committed to reducing social and economic disparities in the developing world, through his field of expertise in water resources. For this reason, Gonzalo has accepted a position at the UNESCO Institute for Water Education (UNESCO-IHE) in Delft, Netherlands.

Permanent email: gespinoza@utexas.edu
This dissertation was typed by the author.